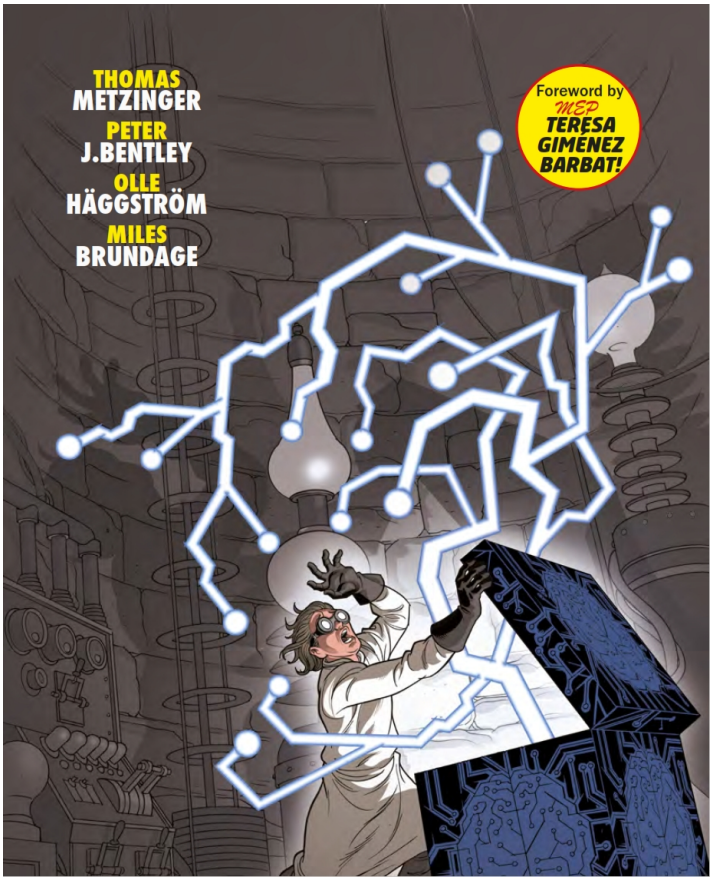
Parlement européen

Faut-il craindre l'intelligence artificielle?



ANALYSE APPROFONDIE

EPRS | Service de recherche du Parlement européen
Unité de la prospective scientifique (STOA)
Mars 2018 - PE 614.547

FR

Faut-il craindre l'intelligence artificielle?

Analyse approfondie

Mars 2018

PE 614.547

AUTEURS

Peter J. Bentley, University College London
Miles Brundage, université d'Oxford
Olle Häggström, université de Chalmers
Thomas Metzinger, université Johannes Gutenberg de Mayence

Avant-propos de María Teresa Giménez Barbat, députée au Parlement européen,
et introduction de Philip Boucher, Unité de la prospective scientifique (STOA)

ADMINISTRATEUR RESPONSABLE STOA

Philip Boucher
Unité de la prospective scientifique (STOA)
Direction de l'Évaluation de l'impact et de la Valeur ajoutée européenne
Direction générale des services de recherche parlementaire
Parlement européen, rue Wiertz 60, 1047 Bruxelles
Courriel: STOA@ep.europa.eu

VERSION LINGUISTIQUE

Original: EN

À PROPOS DE L'ÉDITEUR

Pour contacter la STOA ou pour vous abonner à sa lettre d'information, veuillez écrire à l'adresse suivante: STOA@ep.europa.eu.
Ce document est disponible sur l'internet à l'adresse suivante: http://www.europarl.europa.eu/stoa/

Manuscrit achevé en mars 2018.

CLAUSE DE NON-RESPONSABILITÉ

ISBN 978-92-846-3387-3
doi: 10.2861/587509
QA-01-18-199-FR-N

Table des matières

1. Avant-propos

María Teresa Giménez Barbat, députée au Parlement européen

L'intelligence artificielle (IA) fait depuis plusieurs années une percée remarquable. Une série de programmes visant à tirer le potentiel maximal de la dernière génération de processeurs permet d'obtenir des résultats spectaculaires. L'une des applications les plus remarquables de l'IA est la reconnaissance vocale: alors que les premières versions étaient peu fiables et commettaient de nombreuses erreurs, ces logiciels sont désormais capables de répondre correctement à des demandes très diverses émanant des utilisateurs dans les situations les plus variées. Le domaine de la reconnaissance d'images enregistre aujourd'hui aussi des progrès notables grâce notamment à des programmes capables de reconnaître des visages – et même des chats – sur des vidéos en ligne qui ont été adaptés afin de commander les voitures autonomes qui pulluleront dans nos rues au cours des années à venir. À l'heure actuelle, il est impossible d'imaginer l'avenir en Europe sans une IA avancée, qui aura une incidence sur un nombre toujours plus important d'aspects de notre vie, du travail à la médecine, en passant par l'éducation et les relations interpersonnelles. En février 2017, le Parlement européen a adopté un rapport adressant des recommandations à la Commission européenne en matière de règles de droit civil sur la robotique. Nombreux sont les députés au Parlement européen à avoir entendu, probablement pour la première fois, toute une série d'expressions curieuses telles que les concepts de «robot autonome et intelligent» ou encore de «personnalité électronique».

Si elle veut être véritablement utile, favoriser le progrès et bénéficier au plus grand nombre de citoyens, toute législation future en la matière devra se fonder sur un dialogue avec les experts. Cette préoccupation sous-tend la demande que j'ai adressée au panel d'évaluation des options scientifiques et technologiques (*Scientific and Technology Options Assessment*, STOA) concernant l'organisation d'une manifestation en vue de déterminer si nous devons être optimistes ou pessimistes en ce qui concerne l'IA: pouvons-nous être certains qu'elle profitera à la société?s Nous sommes parvenus à rassembler un panel dirigé par Steven Pinker, professeur de psychologie de Harvard et auteur de publications scientifiques, accompagné de Peter John Bentley, expert en informatique de l'University College de Londres, de Miles Brundage, du Future of Humanity Institute de l'université d'Oxford, d'Olle Häggström, professeur de statistiques mathématiques à l'université de Chalmers et auteur du livre *Here be dragons*, ainsi que du philosophe Thomas Metzinger, de l'université de Mayence, partisan d'un code de déontologie en matière d'IA. Après cette manifestation, ils nous ont tous les quatre envoyé des écrits qui constituent le fondement du présent recueil de textes.

Le lecteur tient en main un recueil d'articles traitant de certaines des idées qui me semblent particulièrement pertinentes pour les responsables politiques et les législateurs. Il est par exemple essentiel de ne pas céder à la tentation de légiférer sur des problèmes qui n'en sont pas. La marche vers une société plus automatisée, dans laquelle la seule intelligence complexe n'est pas humaine, est jalonnée de dangers et de craintes. Les préjugés pessimistes que nous avons hérités de nos ancêtres nous poussent à voir les choses sous un jour plus sombre et à systématiquement nous opposer au progrès technologique. Ils suscitent aussi des craintes excessives, telles que voir une «superintelligence» finir par se retourner contre l'humanité et par créer un avenir «post-humain». Selon Peter Bentley, auteur du texte *Les trois lois de l'intelligence artificielle* , le mythe selon lequel l'IA représente une menace existentielle pour l'humanité est l'un des plus répandus et est source de nombreux malentendus. L'IA repose sur des algorithmes mathématiques qui se contentent de rechercher des modèles. La croyance selon laquelle l'IA pourrait donner naissance à des robots qui chercheraient à dominer le monde n'aurait rien à voir avec la réalité, il ne s'agit là de rien d'autre que science-fiction.

Une autre idée intéressante est que l'IA permettra l'avènement d'une société du bien-être. «Il existe d'innombrables possibilités d'utilisation malveillante de l'IA», explique Miles Brundage, mais si les conditions décrites dans son article *Scaling Up Humanity: The Case for Conditional Optimism about Artificial Intelligence* sont réunies, nous pouvons être très optimistes. L'IA permettra de résoudre des problèmes

complexes et sera désormais responsable de certaines décisions, ce qui évitera ainsi les discriminations et les abus. L'IA va prendre une importance économique considérable au cours des années à venir. Selon une étude réalisée par McKinsey & Co citée par Olle Häggström, la valeur ajoutée économique de l'IA peut, selon des estimations prudentes, être évaluée à 30 milliards de dollars. Thomas Metzinger recense certains des enjeux les plus importants posés par l'avenir de l'IA et formule une série de recommandations pratiques sur la manière dont l'Union européenne pourrait y réagir. Il est certain que nous allons devoir coexister avec l'IA à différents degrés. Nous espérons pouvoir, ensemble, surmonter la plupart de nos peurs et mieux comprendre une technologie qui façonne d'ores et déjà notre avenir.

2. Introduction

Philip Boucher

De manière générale, les êtres humains n'ont jamais vécu aussi longtemps et en aussi bonne santé. Pour beaucoup d'entre nous, ces indicateurs de base suffisent à conclure que nous vivons mieux. Pourtant, il ressort clairement des gros titres de l'actualité qu'une profonde souffrance humaine subsiste. En effet, si l'on pense aux menaces croissantes que font peser le changement climatique, la montée du niveau des océans et les extinctions de masse ou encore aux menaces nucléaires et à l'instabilité politique, il n'y a guère de raisons de se réjouir. Selon les facteurs auxquels la priorité est accordée (égalité, biodiversité, violence, pauvreté, niveaux de CO_2, conflits, appauvrissement de la couche d'ozone) et la façon dont on les mesure, il est tout à fait possible de défendre de manière rationnelle une vision optimiste tout comme pessimiste de l'avenir de l'humanité.

Les avis sont tout aussi partagés à l'égard des nouvelles technologies, comme l'intelligence artificielle (IA), qui devraient avoir une incidence considérable sur l'avenir de l'humanité, pour le meilleur ou pour le pire. L'IA pourrait, pour certains, considérablement améliorer plusieurs aspects de notre vie, comme les prévisions météorologiques ou le diagnostic du cancer. En revanche, d'autres craignent que l'IA ne menace de nombreux emplois et ne rende opaques de nombreux processus de prises de décisions.

Des personnalités connues ont exprimé des positions diamétralement opposées. Elon Musk, par exemple, a dit craindre que l'IA ne fasse peser une menace existentielle sur la race humaine, tandis que Bill Gates estime que la technologie nous rendra plus productifs et créatifs. Au-delà des gros titres cependant, tant Bill Gates qu'Elon Musk reconnaissent que l'IA présente un large éventail de possibilités et d'enjeux, et tous deux appellent à une réflexion sur la façon de gérer son évolution afin d'en tirer le meilleur parti sans néanmoins courir de danger.

Nos espoirs et nos craintes à l'égard de l'IA ne concernent pas seulement un avenir lointain. Ils portent bien souvent sur l'IA d'aujourd'hui, qui a d'ores déjà une influence importante sur nos vies, de toute évidence en bien comme en mal. Pour prendre un exemple, l'IA est à la fois le problème et la solution en matière de fausses informations. Des algorithmes d'IA ont été utilisés pour rendre la justice pénale plus impartiale, mais ont pourtant été accusés de préjugés raciaux

Même si personne ne peut prédire l'évolution future de l'IA, elle devrait sans aucune doute nous offrir de nombreuses possibilités tout en posant de nombreux problèmes, certains plus graves que d'autres. Entre optimisme débridé ou peur paralysante, une position rationnelle unique sur l'avenir de l'IA, s'il devrait y en avoir une, serait certainement plus nuancée. Tant que nous n'en saurons pas davantage sur les incidences de l'IA et sur la capacité de l'humanité à y réagir, il est important de créer des lieux de discussion au sein desquels nous pouvons analyser ces questions, y réfléchir et en discuter et, si nécessaire, y apporter les réponses adéquates. Cette discussion doit rester ouverte à un large éventail de disciplines. La communauté scientifique et technologique a un rôle important à jouer, notamment pour réfléchir aux limites de ce qui est réalisable d'un point de vue technique. Comprendre l'évolution et l'incidence de la technologie dans la société nécessite par exemple des connaissances spécialisées en sciences sociales. Aucune discipline n'a le monopole de la sagesse.

C'est dans ce contexte que le STOA a organisé, le 19 octobre 2017, un atelier au Parlement européen en vue de déterminer s'il est rationnel d'être optimiste à l'égard de l'IA. Steven Pinker (université d'Harvard) a ouvert les débats par une conférence sur le vaste concept d'optimisme rationnel. Ont ensuite pris la parole quatre intervenants de différentes disciplines, Peter J. Bentley, expert en informatique de l'University College London, Miles Brundage, chercheur en politique relative aux technologies de l'université d'Oxford, Olle Häggström, statisticien de l'université Chalmers, et Thomas Metzinger, philosophe de l'université Johannes Gutenberg de Mayence, qui se sont demandés s'il fallait craindre ou non l'IA. Ce débat animé peut encore être visionné en ligne, et nous nous réjouissons du fait que ces quatre intervenants aient accepté d'étayer leurs points de vue dans des

documents de synthèse tous publiés dans le présent recueil. Nous avons donné carte blanche aux auteurs afin qu'ils présentent leurs arguments à leur façon et dans leur style propre afin d'apporter une contribution utile aux débats en cours concernant l'IA au sein de la communauté parlementaire et au-delà. Vu l'attention croissante que les députés européens et les citoyens portent à cette question, de nombreux autres débats et publications suivront dans les années à venir.

3. Les trois lois de l'intelligence artificielle: en finir avec les mythes largement répandus

Peter J. Bentley

Introduction

De nos jours, l'intelligence artificielle (IA) est à la mode. Après plusieurs succès remarqués de nouvelles technologies d'IA, et de nouvelles applications, elle suscite un regain d'intérêt et des experts de nombreuses disciplines s'expriment à son sujet. Chacun y va de son grain de sel, qu'il s'agisse du commun des mortels, de responsables politiques, de philosophes, d'entrepreneurs ou de lobbyistes professionnels. Ces avis font toutefois rarement appel aux personnes qui maîtrisent le mieux l'IA: les informaticiens et ingénieurs qui passent leurs journées à concevoir des solutions intelligentes, à les appliquer à de nouveaux produits et à les tester. Cet article présente le point de vue d'un informaticien expérimenté dans la création de technologies intelligentes, le but étant d'assurer un certain équilibre et d'apporter un avis éclairé sur la question.

Démythifier

L'un des points de vue les plus incroyables souvent répétés est que l'IA représente un danger pour l'humanité, voire même une «menace existentielle». Certains affirment que l'IA pourrait se développer spontanément et violemment, comme une espèce de cancer à l'intelligence exponentielle. Même si nous commençons par quelque chose de simple, cette intelligence s'améliorerait sans que nous puissions la contrôler. Et du jour au lendemain, toute l'espèce humaine se retrouverait à lutter pour sa survie (2015).

Cette perspective est absolument terrifiante, ce qui explique que tant de films de science-fiction s'en inspirent. Néanmoins, malgré les propos tenus par de fervents observateurs, philosophes et autres personnes qui devraient avoir l'intelligence de ne pas échafauder de tels scénarios, il s'agit là d'un pur fantasme. La réalité est tout autre: L'IA, comme toute forme d'intelligence, ne peut évoluer que lentement et péniblement. Il n'est pas facile de devenir intelligent.

Deux types d'IA ont toujours coexisté: celle du monde réel et celle de la fiction. L'IA du monde réel est celle qui nous entoure: la reconnaissance vocale de Siri ou de l'Echo, les systèmes cachés de détection des fraudes de nos banques ou encore les systèmes de lecture des plaques minéralogiques utilisés par la police (Aron, 2011 Siegel, 2013; Anagnostopoulos, 2014). La réalité de l'IA est que nous mettons au point des centaines de logiciels intelligents différents et hautement spécialisés pour résoudre un million de problèmes différents dans différents produits. Il en va ainsi depuis la naissance même de l'IA, qui coïncide avec la création des premiers ordinateurs (Bentley, 2012). Les technologies de l'IA sont déjà intégrées aux logiciels et aux machines qui nous entourent. Elles ne constituent toutefois que des formes de technologie astucieuse; elles sont l'équivalent informatique des rouages et ressorts des machines mécaniques. Et, comme dans le cas d'un engrenage cassé ou d'un ressort qui saute, leur défaillance peut entraîner la panne du produit concerné. Tout comme un rouage ou un ressort ne se transforme pas comme par magie en un robot tueur, nos logiciels intelligents intégrés à leurs produits ne peuvent pas se transformer en une IA malveillante.

L'IA du monde réel sauve des vies en déclenchant des mécanismes de sécurité (freinage automatique des voitures, voire même véhicules autonomes). L'IA du monde réel nous aide à optimiser les processus ou à prédire les pannes, à améliorer l'efficacité ou à réduire les déchets nocifs pour l'environnement. La seule raison pour laquelle des centaines d'entreprises sont spécialisées dans l'IA et des milliers de chercheurs et d'ingénieurs étudient ce domaine est qu'ils aspirent à trouver des solutions qui aident les individus et améliorent leur quotidien (Richardson, 2017).

L'autre type d'IA, y compris cette IA générale super-intelligente qui va tous nous exterminer, n'est que fiction. Les chercheurs travaillent en général sur le premier type d'IA. Néanmoins, puisque cet article se doit de rétablir un équilibre en faveur d'un sens commun rationnel, les paragraphes qui suivent s'efforcent de casser certains mythes dans ce domaine. Dans cet article, je présente les «trois lois de l'IA» afin d'expliquer pourquoi ces mythes sont fantaisistes, pour ne pas dire ridicules. Ces «lois» sont une simple synthèse de plusieurs dizaines d'années de recherches scientifiques en IA, simplifiées pour les présenter au lecteur non-initié.

Mythe 1: une IA auto-modifiée se rendra super-intelligente

Certains observateurs pensent qu'il existe un risque que l'IA «brise ses chaînes» et se rende elle-même super-intelligente (Häggström, 2016).

La première loi de l'IA nous explique la raison pour laquelle ce scénario est impossible.

Première loi de l'IA: la difficulté engendre l'intelligence

Nos recherches dans le domaine de la vie artificielle nous révèlent que l'intelligence n'existe que pour répondre à des besoins urgents. En l'absence du bon type de problèmes à résoudre, l'intelligence ne peut se faire jour ni progresser (Taylor *et al., 2014).* L'intelligence est nécessaire uniquement si ces problèmes sont variés et imprévisibles. L'intelligence n'évoluera pour résoudre ces problèmes que si son avenir dépend de sa réussite.

Pour créer une IA simple, nous mettons au point un algorithme visant à résoudre un problème particulier. Pour transformer cette IA en une IA générale, nous devons présenter à notre IA en développement des problèmes de complexité et de variété croissantes, élaborer de nouveaux algorithmes pour les résoudre et garder les algorithmes qui ont apporté une solution. Faute de nouveaux défis à résoudre en permanence, et sans récompense en cas de réussite, nos IA ne gagneront pas un point de QI supplémentaire.

Les chercheurs en IA ne le savent que trop bien. Un robot capable d'accomplir efficacement une tâche n'augmentera jamais ses capacités si nous ne l'obligeons pas à progresser (Vargas *et al.*, 2014). Prenons un exemple: le système de reconnaissance automatique des plaques minéralogiques utilisé par la police est une forme spécialisée d'IA conçue pour résoudre un problème bien précis, à savoir lire des plaques minéralogiques. Même si nous ajoutions certains processus à cette IA pour lui permettre de se modifier elle-même, elle n'accroîtrait jamais son intelligence si elle n'est pas amenée à faire face à un problème nouveau et complexe. Sans besoin urgent, l'intelligence n'est qu'une perte de temps et d'énergie. Le monde naturel en offre de nombreuses illustrations: la plupart des problèmes que présente la nature ne nécessitent pas de cerveaux pour les résoudre. Rares sont les organismes qui ont dû déployer les efforts considérables nécessaires pour se doter de cerveaux, et plus rares encore ceux qui développent des cerveaux extrêmement complexes.

La première loi de l'IA nous explique que l'intelligence artificielle est un objectif extrêmement difficile à atteindre, qui nécessite des efforts considérables et des conditions parfaites. Il n'y aura pas de fuite d'IA, ni d'IA se développant de manière automome hors de notre contrôle. Il n'y aura pas de singularités. L'IA n'ira jamais plus loin que le niveau d'intelligence que nous l'encourageons (ou la forçons) à acquérir, sous la contrainte.

Par ailleurs, même si nous pouvions créer une super-intelligence, rien n'indique que cette IA super-intelligente nous voudrait du mal. Ces affirmations sont profondément erronées et s'inspirent sans doute du comportement humain, qui est effectivement très violent. Les IA ne disposeront toutefois pas d'une intelligence humaine. Il est presque certain que notre véritable avenir s'inscrira dans le droit fil de notre situation présente: les IA évolueront avec nous et seront conçues pour répondre à nos besoins, tout comme nous avons fait évoluer nos cultures, notre bétail et nos animaux de compagnie pour répondre à nos besoins (Thrall *et al.*, 2010). Nos chiens et nos chats n'ont pas pour projet de tuer tous les

humains. De même, une IA plus avancée nous correspondra si étroitement qu'elle sera intégrée au sein de nos vies et de nos sociétés. Elle ne souhaitera pas davantage nous tuer tous que se tuer elle-même.

Mythe 2: moyennant des ressources suffisantes (neurones/ordinateurs/mémoire), une IA sera plus intelligente que les humains

Certains observateurs affirment que «plus est synonyme de mieux». Si un cerveau humain compte cent milliards de neurones, une IA comptant mille milliards de neurones simulés sera plus intelligente qu'un être humain. Si un cerveau humain est l'équivalent de tous les ordinateurs sur l'internet, une IA lâchée sur l'internet possèdera une intelligence humaine. En fait ce n'est pas la quantité qui compte, mais l'organisation de ces ressources, comme l'explique la deuxième loi de l'IA.

Deuxième loi de l'IA: l'intelligence nécessite une structure adéquate

Les cerveaux peuvent être structurés de manières très différentes. Chaque type de problème nécessite un nouveau modèle pour le résoudre. Pour comprendre ce que nous voyons, nous avons besoin d'un type particulier de structure neuronale qui diffère de celui requis pour faire bouger nos muscles ou encore pour mémoriser des souvenirs. La biologie nous enseigne qu'il ne faut pas beaucoup de neurones pour être incroyablement intelligent. L'astuce consiste à organiser correctement ces neurones, à écrire l'algorithme optimal pour chaque problème (Gardner et Mayford, 2012).

Pourquoi ne pas utiliser les mathématiques pour créer des IA?

Nous utilisons beaucoup de mathématiques intelligentes, ce qui explique que certaines méthodes d'apprentissage des machines produisent des résultats prévisibles qui nous permettent de comprendre précisément ce que les IA peuvent et ne peuvent pas faire. La plupart des solutions pratiques sont toutefois imprévisibles car elles sont trop complexes et leurs algorithmes ont recours à l'aléatoire, de sorte que nos mathématiques sont dépassées, et car elles reçoivent souvent des données entrantes imprévisibles. Nous ne disposons pas d'outils mathématiques permettant de prédire les capacités d'une nouvelle IA, mais nous en avons d'autres qui nous indiquent les limites des calculs. Alan Turing a participé à l'invention de l'informatique en mettant au jour une limite particulière: il est impossible de prédire si un algorithme arbitraire (y compris une IA) arrêtera un jour ses calculs ou non (Turing, 1937). Selon un autre théorème, celui du «rien n'est gratuit», les performances d'un algorithme n'excèderont jamais celles des autres pour l'ensemble des problèmes; en d'autres termes, un nouvel algorithme adapté à chaque nouveau problème est nécessaire si nous souhaitons disposer de l'intelligence la plus efficace (Wolpert, 1996; Wolpert et Macready, 1997). Le théorème de Rice nous enseigne quant à lui qu'il est impossible qu'un algorithme corrige parfaitement un autre algorithme, et donc que même si une IA peut évoluer par elle-même, elle ne sera jamais en mesure de déterminer si une modification fonctionne dans tous les cas sans essais empiriques (Rice, 1953).

Pour créer une IA, nous devons concevoir de nouvelles structures/de nouveaux algorithmes spécialisés pour chaque problème que cette IA devra résoudre. Des types de problèmes différents nécessitent des structures différentes. Un problème rencontré pour la première fois peut nécessiter le développement d'une nouvelle structure entièrement nouvelle. Il n'existe aucune structure universelle qui soit adaptée à tous les problèmes – c'est ce que nous enseigne le théorème du «rien n'est gratuit» (Wolpert, 1996; Wolpert et Macready, 1997) (voir encadré). Dès lors, la création d'une intelligence toujours plus avancée, ou la capacité à résoudre des problèmes toujours plus variés, est un processus d'innovation permanent qui nécessite l'invention de nouvelles structures adaptées à chaque nouvelle difficulté. L'un des grands problèmes de la recherche en IA consiste à déterminer les structures ou les algorithmes qui permettent de résoudre chacun de ces problèmes. La recherche en est encore à ses premiers balbutiements dans ce domaine, c'est pourquoi les IA actuelles ne sont encore dotées que d'une intelligence extrêmement limitée.

À mesure que nous rendons nos IA plus intelligentes (ou si nous trouvons un jour le moyen de créer des IA capables d'évoluer par elles-mêmes à l'infini), nous rencontrons encore de nouveaux problèmes.

Il est impossible de concevoir l'intelligence d'un seul coup: nous ne possédons pas d'outils mathématiques permettant de prédire les capacités d'une nouvelle structure et nous n'avons pas de compréhension siffisante de la correspondance entre différents algorithmes/structures et différents problèmes. Notre seule possibilité de créer des intelligences plus avancées consiste à progresser pas à pas en tâtonnant.

Nous devons intégrer chaque nouvelle structure à l'intelligence existante sans perturber les structures déjà en place. Il s'agit d'une tâche extrêmement difficile à réaliser qui peut résulter en une superposition de couches de nouvelles structures, chacune d'entre elle fonctionnant à l'unisson avec les structures antérieures, comme on l'observe dans le cerveau humain. Si nous souhaitons un cerveau plus intelligent encore, comme le nôtre, nous pouvons également ajouter la capacité de certaines structures à se réorienter lorsque d'autres structures sont endommagées, et donc évoluer par elles-mêmes jusqu'à ce qu'elles soient en mesure d'assurer au moins partiellement les fonctions perdues. Nous ne savons pas vraiment non plus comment y parvenir.

La deuxième loi de l'IA nous explique que les ressources ne suffisent pas. Nous devons encore concevoir de nouveaux algorithmes et de nouvelles structures au sein des IA (et à l'appui de celles-ci) pour chaque nouveau problème qu'elles sont amenées à résoudre.

C'est pourquoi il est impossible de créer des intelligences à usage général à l'aide d'une approche unique. Aucune IA au monde (pas même l'«apprentissage profond» dont on parle tant) ne peut utiliser la même méthode pour comprendre la langue, conduire une voiture, apprendre à jouer à un jeu vidéo complexe, faire courir un robot dans une rue encombrée, laver la vaisselle dans l'évier ou planifier la stratégie d'investissement d'une entreprise. Lorsque le cerveau humain accomplit de telles tâches, il utilise une multitude de structures neuronales différentes combinées de diverses façons, chacune étant conçue pour résoudre un sous-problème précis. Nous ne sommes pas en mesure de créer de tels cerveaux; nous construisons donc une solution intelligente spécialisée pour chaque problème et les utilisons isolément.

Mythe 3: la vitesse des ordinateurs doublant tous les 18 mois, les IA vont utiliser cette puissance de calcul pour progresser de manière exponentielle

Certains observateurs affirment que la rapidité brute de calcul permettra de surmonter toutes les difficultés rencontrées en matière de création d'IA. Pour peu qu'elle utilise des ordinateurs suffisamment rapides, une IA parviendra à apprendre et à nous dépasser dans la réflexion. Puisque la vitesse des processeurs double tous les 18 mois environs depuis des décennies, cette issue est forcément inévitable. Malheureusement, ce point de vue ne tient pas compte de l'incidence d'un autre phénomène exponentiel qui freine fortement le développement des IA, à savoir les essais.

Troisième loi de l'IA: l'intelligence nécessite des essais poussés

Une intelligence supérieure nécessite les conceptions les plus complexes de l'univers. Toutefois, la moindre modification apportée en vue d'améliorer la conception d'une intelligence est susceptible de détruire une partie voire l'ensemble de ses capacités existantes. Ce problème est d'autant plus délicat que nous ne possédons pas d'outils mathématiques capables de prédire les capacités d'une intelligence générale (voir encadré). C'est pourquoi chaque nouveau modèle d'intelligence nécessite des essais poussés de tous les problèmes qu'il est censé résoudre. Des essais partiels ne suffisent pas: il faut tester l'intelligence dans toutes les permutations possibles du problème pour sa durée de vie prévue, faute de quoi ses capacités ne seront pas fiables.

Tous les chercheurs en IA ne connaissent que très bien cette dure vérité: pour créer une IA, il est nécessaire de l'entraîner et de tester de manière poussée toutes ses capacités dans son environnement prévu à chaque étape de sa conception. Comme le disait Marvin Minsky, fondateur de l'IA, «[...] on raconte tellement d'histoires sur la façon dont les choses pourraient mal tourner, mais je ne parviens pas à les prendre au sérieux car j'ai du mal à imaginer que quelqu'un puisse les installer à grande échelle

sans avoir réalisé des essais suffisants» (Achenbach, 2016). Plus que tout autre aspect, ce sont les essais qui nécessitent le plus de temps. Cette contrainte de temps constitue un frein à la conception d'intelligences. Dans le pire des cas, le niveau d'essais requis augmente de manière exponentielle pour chaque gain d'intelligence différentiel.

Pour le comprendre, imaginons une intelligence qui est capable de reconnaître dix couleurs différentes et qui doit distinguer deux types d'éléments sur la base d'une seule caractéristique, à savoir la couleur. Dans ce cas, l'IA ne peut distinguer qu'un nombre maximal de dix différents types d'éléments et les classer en deux catégories. Si ses capacités sont améliorées de manière à traiter deux caractéristiques (la couleur et dix nuances de luminosité, par exemple), elle pourra reconnaître un nombre maximal de 100 types d'éléments. De même, si elle pouvait traiter 100 caractéristiques, elle pourrait comprendre 10 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 éléments. Comment pouvons-nous savoir que cette IA les classera tous correctement? Cette question est d'autant plus pertinente pour les applications critiques pour la sécurité. Imaginons que cette IA conduit votre voiture et que les deux catégories d'éléments sont «route» et «obstacle».

Or, la réalité est bien plus complexe. De nombreuses intelligences naturelles sont capables de distinguer un nombre inimaginable d'éléments sur la base de milliers ou de millions de caractéristiques, chacune étant susceptible de représenter des milliers ou des millions de valeurs. Tester ces intelligences prend évidemment énormément de temps, ce qui n'y a rien d'étonnant. L'évolution crée ses cerveaux biologiques en testant des milliers de milliards de cerveaux simultanément dans chaque scénario possible pendant des milliards d'années, et seuls les cerveaux les plus performants peuvent contribuer à la conception des futurs cerveaux. Il s'agit d'une méthode d'essai hautement efficiente et efficace dont nous ne pouvons que rêver.

Cet énorme besoin de tester les IA a des implications d'une portée considérable. Il est impossible de concevoir de meilleures IA sans les tester à chaque étape. Google a procédé à des essais de ses véhicules autonomes pendant des années et continue de le faire (en juin 2016, Google avait réalisé au total 2 777 585 km d'essais de son parc de véhicules en mode autonome) (Google, 2016). Tous les ingénieurs et chercheurs dans le domaine de l'IA savent qu'il est impossible d'utiliser l'IA dans une application critique pour la sécurité sans avoir procédé à des essais suffisants. Dans un avenir proche, des certifications permettant de savoir précisément comment une IA accomplit certaines tâches bien définies seront requises. Et si l'IA poursuit son apprentissage ou est mise à jour, nous ne pouvons pas partir du principe qu'elle reste en mesure de réussir un essai réussi précédemment. Tout comme un pilote humain, une IA qui poursuit son apprentissage doit constamment subir de nouveaux tests afin de conserver sa certification[1].

La troisième loi de l'IA nous apprend qu'à mesure que l'intelligence progresse, le temps nécessaire à la tester est susceptible d'augmenter de manière exponentielle. À terme, ces essais peuvent imposer des limites pratiques à l'intelligence artificielle réalisable et fiable. Tout comme il devient de plus en plus difficile d'accélérer à mesure que l'on approche de la vitesse de la lumière, il devient de plus en plus difficile d'améliorer l'intelligence à mesure que nous créons des cerveaux plus intelligents.

[1] Par ailleurs, toute interaction entre un humain et des IA nécessite également que nous suivions une formation et obtenions une nouvelle certification. Une automobile dotée d'une IA qui prend le contrôle dans certaines circonstances devient un fardeau lorsque cette IA atteint les limites de ses capacités et que le conducteur n'a pas été formé à demeurer suffisamment vigilant pour reprendre les commandes (Eriksson et Stanton, 2017).

Une fois encore, il s'agit d'une raison fondamentale pour laquelle les recherches et applications en matière d'intelligence artificielle s'efforcent de trouver des solutions intelligentes à des problèmes bien précis[2].

Conclusions

L'IA a vu le jour avec la naissance des ordinateurs. Elle existe depuis suffisamment longtemps pour avoir connu des périodes d'euphorie poussant certains experts reconnus à faire des déclarations à peine croyables (Bentley, 2012). Claude Shannon fut l'un des plus grands pionniers de l'informatique et de l'IA. En 1961, il affirmait: "«Je suis convaincu que, d'ici dix ou quinze ans, nos laboratoires produiront quelque chose ressemblant de fort près aux robots de science fiction.»[3] Il pensait qu'au milieu des années 1970, les machines marcheraient, parleraient et réfléchiraient de manière autonome. Néanmoins, quarante ans plus tard, c'est à peine si nous parvenons à faire marcher un robot, et, évidemment, aucun d'entre eux ne pense seul. Aujourd'hui, certaines enquêtes (rassemblant une diversité de points de vue) prédisent qu'il existe une «probabilité de 50 % que l'IA dépasse l'être humain dans tous les domaines d'ici à 45 ans» (Grace *et al*., 2017). Tout cela, nous l'avons déjà entendu des centaines de fois. Et cette prévision s'avérera pourtant tout aussi erronée.

Ne croyez pas tout le battage médiatique. Nous sommes bien incapables de prédire l'avenir et les prévisions (même celles des experts de renommée mondiale) sont presque toujours complètement inexactes. En fin de compte, l'histoire nous apprend que ce battage médiatique est a l'origine des périodes de récession qu'observe la recherche en IA (Bentley, 2012). Les avancées spectaculaires engendrent de la publicité à grande échelle, donc des investissements importants et une vague de réglementation. Et puis, inévitablement, tout le monde redescend sur terre. L'IA n'est pas à la hauteur de tout le battage médiatique qu'elle suscite. Les investissements se tarissent. La réglementation étouffe l'innovation. Et «IA» devient un gros mot que plus personne n'ose prononcer. Un nouvel hiver de l'IA détruit les progrès accomplis.

Les propos alarmistes et les prévisions ridicules n'ont pas leur place dans le progrès scientifique ni dans l'élaboration des politiques, qu'ils restent cantonnés aux salles de cinéma. Une discussion posée et rationnelle est par contre de la plus haute importance. L'IA est aujourd'hui utilisée dans de nouvelles applications critiques pour la sécurité. Les propos alarmistes détournant l'attention de son object principal pourraient coûter des vies. Au lieu de nous focaliser sur ce qui pourrait arriver si un scénario de science-fiction devenait réalité, nous ferions mieux de nous concentrer sur une nouvelle réglementation en matière de sécurité et sur une certification pour chaque application de l'IA critique pour la sécurité. Qu'en est-il des nouveaux essais de sécurité routière et des certifications pour voitures sans chauffeur? Des nouveaux examens de permis de conduire pour les conducteurs humains qui possèdent des voitures sans chauffeur? Des nouveaux témoins lumineux agréés informant les piétons que la voiture les a vus et qu'ils peuvent traverser la route en toute sécurité? Ou encore des réglementations empêchant les services de presse conservateurs en matière d'IA de polariser davantage encore le débat au sein de la population? (Cesa-Bianchi *et al*., 2017) L'heure est venue de laisser ces inepties de côté et de se concentrer désormais sans plus attendre sur la réalité. Comment assurer aujourd'hui la sécurité de chaque nouvelle application d'un logiciel intelligent?

---

[2] Une autre raison fondamentale réside dans nos propres cerveaux: à l'heure actuelle, et dans un avenir prévisible, nous ne sommes pas suffisamment intelligents pour créer de l'intelligence. Nous ne comprenons pas le fonctionnement des cerveaux biologiques. Nous ne savons pas pourquoi certaines de nos meilleures méthodes d'IA fonctionnent. Nous ne savons pas comment les améliorer. Notre propre ignorance constitue un frein considérable au progrès.

[3] Extrait d'une interview de Claude Shannon lors de l'émission télévisée *The Thinking Machine*, dans la série de documentaires «Tomorrow», 1961. Copyright CBS News.

L'intelligence artificielle recèle un potentiel fascinant nous permettant d'améliorer notre quotidien, de nous aider à vivre plus heureux et en meilleure santé et de créer de nombreux nouveaux emplois. La création d'IA englobe nombre des plus grands exploits scientifiques et techniques de tous les temps. Il s'agit d'une nouvelle révolution technologique. Cette révolution ne se fera toutefois pas comme par magie. Les trois lois de l'IA nous enseignent que si nous voulons créer des IA plus avancées, nous devons soumettre progressivement des problèmes plus ardus à nos IA, concevoir méticuleusement de nouvelles structures intelligentes pour leur permettre de résoudre ces problèmes et procéder à des essais de grande envergure pour confirmer que nous pouvons leur faire confiance en ce qui concerne la résolution de ces problèmes. Des milliers de scientifiques et d'ingénieurs compétents suivent précisément et sans relâche ces étapes (problème, hypothèse de solution, essais) pour nous apporter la moindre amélioration, aussi petite soit-elle, car tels sont notre processus de conception et notre méthode scientifique. Ne craignez pas l'IA, admirez plutôt l'ardeur et les compétences de ces experts humains qui consacrent leur vie à contribuer à sa création. Et gardez à l'esprit que l'IA contribue chaque jour à améliorer notrequotidien.

Références

Achenbach, J. (2016). Professeur Marvin Minsky: mathématicien et inventeur inspiré par Alan Turing, aujourd'hui l'un des pionniers de l'intelligence artificielle. Rubrique nécrologique, *The Independent*, 29 janvier 2016.

Anagnostopoulos, C-N E., (2014) License Plate Recognition: A Brief Tutorial. IEEE Intelligent Transportation Systems Magazine. Volume: 6, numéro: 1, pp. 59–67. DOI: 10.1109/MITS.2013.2292652

Aron, J. (2011) How innovative is Apple's new voice assistant, Siri?. New Scientist Vol 212, Numéro 2836, 29 octobre 2011, p. 24. https://doi.org/10.1016/S0262-4079(11)62647-X

Barrat, J. (2015) Why Stephen Hawking and Bill Gates Are Terrified of Artificial Intelligence. Huffington Post.

Bentley, P. J. (2012) Digitized: The science of computers and how it shapes our world. OUP Oxford. ISBN-13: 978-0199693795.

Cesa-Bianchi, N., Pontil, M., Shawe-Taylor, J., Watkins, C., et Yilmaz, E. (2017) Proceedings of workshop: Prioritise me! Side-effects of online content delivery. The problem of bubbles and echo-chambers: new approaches to content prioritisation for on-line media, 12 juin 2017. Knowledge 4 All Foundation. Londres, Royaume-Uni.

Eriksson, A., & Stanton, N. A. (2017). Driving performance after self-regulated control transitions in highly automated vehicles. Human Factors: The Journal of the Human Factors and Ergonomics Society. DOI: 10.1177/0018720817728774

Garner, A. et Mayford, M. (2012) New approaches to neural circuits in behaviour. Learn. Mem. 2012. 19: 385-390. Doi: 10.1101/lm.025049.111

Google (2016). «Google Self-Driving Car Project Monthly Report - June 2016» (PDF). Google. Téléchargé le 15 juillet 2016. https://static.googleusercontent.com/media/www.google.com/en//selfdrivingcar/files/reports/report-0616.pdf

Grace, K., Salvatier, J. Dafoe,A., Zhang, B. et Evans, O. When Will AI Exceed Human Performance? Evidence from AI Experts. arXiv:1705.08807

Häggström, O. (2016) Here Be Dragons: Science, Technology and the Future of Humanity. OUP Oxford.

Rice, H. G. (1953). Classes of Recursively Enumerable Sets and Their Decision Problems. Trans. Amer. Math. Soc. 74, 358-366.

Richardson, J. (2017) Three Ways Artificial Intelligence is Good for Society. IQ magazine, Intel. Disponible en ligne: https://iq.intel.com/artificial-intelligence-is-good-for-society/

Siegel, E. (2013) Predictive Analytics: The Power to Predict who will Click, Buy, Lie or Die. John Wiley & Sons, Inc. ISBN: 978-1-118-35685-2.

Taylor, T., Dorin, A., Korb, K. (2014) Digital Genesis: Computers, Evolution and Artificial Life. Article présenté lors de la 7e conférence de philosophie des sciences de Munich-Sydney-Tilburg: Evolutionary Thinking, University of Sydney, 20-22 mars 2014. arXiv:1512.02100 [cs.NE]

Thrall, P. H., Bever, J. D., et Burdon, J. J. (2010) Evolutionary change in agriculture: the past, present and future. Evol Appl.3(5-6): 405–408. doi: 10.1111/j.1752-4571.2010.00155.x

Turing, A. (1937) On computable numbers, with an application to the Entscheidungsproblem, Proceedings of the London Mathematical Society, Series 2, Volume 42, pp 230–265, doi:10.1112/plms/s2-42.1.230

Vargas, P. A., Di Paolo, E. A., Harvey, I. et Husbands, P. (Eds) (2014) The Horizons of Evolutionary Robotics. MIT Press.

Wolpert, D.H. (1996). The Lack of A Priori Distinctions between Learning Algorithms. Neural Computation, pp. 1341-1390.

Wolpert, D.H., Macready, W.G. (1997). No Free Lunch Theorems for Optimization. IEEE Transactions on Evolutionary Computation 1, 67.

4. Montée en puissance de l'humanité: les raisons d'un optimisme conditionnel à l'égard de l'intelligence artificielle

Miles Brundage

Introduction

Les avis des experts concernant la chronologie des évolutions futures en matière d'intelligence artificielle (IA) varient considérablement. Certains prévoient l'avènement d'une IA équivalente à l'intelligence humaine d'ici quelques décennies, tandis que d'autres ne l'envisagent que dans un avenir nettement plus lointain (Grace *et al.*, 2017). De même, pour certains, l'évolution de l'IA sera bénéfique et pour d'autres, néfaste pour la civilisation humaine; d'aucuns sont convaincus que l'IA sera extrêmement bénéfique, d'autres s'attendant à ce qu'elle soit extrêmement néfaste (au point de menacer l'humanité d'extinction), et bien d'autres encore ont des avis plus partagés (AI Impacts, 2017). Bien que les risques liés au développement de l'IA aient suscité un grand intérêt ces derniers temps (Bostrom, 2014; Amodei et Olah *et al.*, 2016), il n'y a eu que peu de discussions systématiques sur les avantages que l'IA pourrait plus précisément nous apporter à long terme.

Dans le présent article, je ne tente pas de définir l'évolution probable mais plutôt de plaider en faveur d'un *optimisme conditionnel* à l'égard de l'IA et de préciser les raisons pour lesquelles l'IA pourrait devenir une technologie transformatrice pour l'humanité, et ce dans son intérêt. J'entends par là que, si l'humanité parvient à relever les défis techniques, éthiques et politiques liés au développement et à la diffusion de technologies puissantes d'IA, cette dernière pourrait avoir une incidence énorme et potentiellement très bénéfique sur le bien-être de l'humanité. Pour justifier cette conclusion, je commencerai par passer en revue les caractéristiques de l'IA susceptibles d'avoir une incidence (positive ou négative) considérable sur le bien-être de l'humanité à long terme. Ensuite, je décrirai brièvement certaines conditions nécessaires à sa réussite, c'est-à-dire les défis à relever pour parvenir à l'avenir radieux que les caractéristiques de l'IA rendent possible. Ensuite, dans le corps de l'article, j'énumèrerai trois raisons distinctes pour lesquelles il convient de s'attendre (sous certaines conditions) à une incidence positive considérable de l'IA sur l'humanité: une IA puissante accélérerait considérablement la réalisation de tâches (*accélération des tâches*), permettrait une coordination plus efficace et à plus grande échelle des individus et des institutions (*amélioration de la coordination*) et permettrait de réorienter la vie des êtres humains vers la réalisation d'objectifs qu'ils jugent intrinsèquement épanouissants, tout en préservant un niveau de vie élevé sans devoir d'accomplir un travail non désiré (*société des loisirs*).

Aucun de ces résultats ne garantit l'évolution favorable de l'IA, mais je pars de l'hypothèse que l'IA est indispensable pour parvenir à chacun d'eux. Pour certains, l'évolution de l'IA est tellement risquée qu'il vaut mieux l'éviter, mais j'avance pour ma part plutôt que l'IA constituera un élément essentiel de la prospérité de l'espèce humaine à long terme et conclus sur une vision positive de ce à quoi pourrait ressembler le résultat final.

Caractéristiques de l'intelligence artificielle

L'IA désigne un ensemble de recherches et d'ingénierie axées sur l'utilisation des technologies numériques pour créer des systèmes capables de réaliser certaines tâches (souvent à l'issue d'un apprentissage) dont il est généralement admis que leur réalisation par un être humain ou un animal nécessite de l'intelligence. Ce domaine a progressé très rapidement ces dernières années après des décennies de résultats insatisfaisants. Au rang des récents succès notables de l'IA figure le dépassement des capacités humaines au jeu de Go et dans une série de tâches de traitement de l'image. Les technologies d'IA sont très présentes dans la vie moderne, notamment dans des applications

couramment utilisées, telles que les moteurs de recherche, la reconnaissance vocale sur les téléphones et la traduction automatique en ligne.

Élément plus important encore que toute réussite particulière de l'IA dans le cadre d'une tâche bien précise, l'IA associe les caractéristiques des technologies numériques en général (notamment l'*évolutivité* par la reproduction des programmes et l'accélération de leur exécution) à des caractéristiques généralement considérées comme propres à l'homme (la *compétence*). L'importance de l'IA réside donc dans une large mesure dans sa capacité à améliorer la performance de tâches intelligentes, par exemple lorsqu'un système de traduction automatique permet la traduction de textes par des millions d'utilisateurs simultanément. Outre cette caractéristique de *compétence évolutive*, les IA peuvent en principe s'atteler à presque n'importe quelle tâche (Bostrom, 2014), ce qui représente à la fois un risque et une chance. Enfin, que ce soit dans des domaines précis aujourd'hui ou plus largement dans le cadre de prises de décisions à long terme, l'IA peut dépasser les performances humaines, ce qui permettra de consacrer un nombre important de systèmes rapides et compétents à la réalisation de presque n'importe quelle tâche. Ce sont ces caractéristiques de l'IA qui guident la réflexion relative à ses conséquences sociales développée ci-dessous.

Conditions de la réussite

Une technologie puissante et souple aura un large éventail de conséquences sociales (comme c'est le cas par exemple de l'électricité). Toutefois, contrairement à l'électricité, les systèmes d'IA ont le potentiel d'assurer des fonctions nettement plus diverses aujourd'hui, et plus encore à l'avenir. L'IA est susceptible de faire l'objet de nombreuses utilisations malveillantes (Brundage et Avin *et al.*, 2018), et nombreuses sont les manières de l'utiliser de manière néfaste sans en avoir l'intention, comme dans le cas du parti-pris algorithmique (Kirkpatrick, 2016). Pour en tirer les avantages décrits ci-dessous à long terme, il faudra chercher à éviter de nombreuses conséquences néfastes. La plus grande difficulté sera sans doute de régler le *problème du contrôle*: nous devrons apprendre à faire en sorte que les systèmes d'IA atteignent les objectifs recherchés (Bostrom, 2014; Amodei et Olah *et al.*, 2016; Bostrom, Dafoe, et Flynn, 2017) sans causer de préjudice pendant leur processus d'apprentissage, sans interpréter de manière erronée ce que l'on attend d'eux et sans résister au contrôle par l'être humain. Bien que les systèmes d'IA actuels ne disposent que de capacités limitées par rapport à l'être humain et que certaines craintes excessives (par exemple le risque de voir un système d'IA résister à sa mise hors fonction) soient très peu vraisemblables, la résolution du problème du contrôle constituera une condition préalable essentielle à long terme pour permettre à des systèmes d'IA plus puissants d'avoir une incidence favorable sur la société. Il faudra en outre parvenir à relever les défis politiques que pose l'IA, notamment les risques associés à une concentration excessive de pouvoir et de richesse (Bostrom, Dafoe et Flynn, 2017) et les courses au développement risquées qui poussent à négliger la sécurité pour en tirer un avantage (Armstrong *et al.*, 2016; Bostrom, 2017).

Pour recentrer le débat, je pars ci-dessous de l'hypothèse que ces défis sont relevés avec succès et je m'étends sur les différentes façons dont le résultat pourrait s'avérer extrêmement bénéfique. Comme indiqué ci-dessus, ce choix ne doit pas être interprété comme une prévision mais comme un exercice visant à examiner plus en détail un aspect du rapport coûts/avantages. Après avoir présenté chacun des motifs d'optimisme, j'associerai tous ces motifs dans une vision positive globale d'un avenir possible avec une IA plus avancée.

Motifs d'optimisme: accélération des tâches, amélioration de la coordination et société des loisirs

Accélération des tâches

La compétence évolutive de l'IA favorise l'exécution d'un grand nombre de tâches plus rapidement que ne le permettrait une autre approche, qu'il s'agisse de tâches accessibles aux êtres humains (moyennant suffisamment de temps et de ressources) ou de tâches que l'être humain est incapable de réaliser en

raison de ses limites cognitives ou organisationnelles. Les systèmes d'IA ont déjà affiché des performances de niveau humain et des performances dépassant celles de l'homme. Au jeu de Go, par exemple, l'IA a dépassé l'humain au cours de ces quelques dernières années. Les systèmes technologiques de niveau humain (voire inférieurs au niveau de performance humain) sont utilisés de manières très diverses, comme pour l'exécution de tâches fastidieuses ou chronophages. La traduction automatique en est un excellent exemple: chaque amélioration des systèmes de traduction automatique, aussi petite soit-elle, peut être appliquée relativement rapidement à un large éventail de paires de langues et à des millions d'utilisateurs. De même, la reconnaissance vocale n'atteint pas encore la performance humaine en toutes circonstances mais permet souvent un gain de temps pour les utilisateurs d'appareils électroniques qui préfèrent ne pas avoir à taper chaque mot.

L'*accélération des tâches* propre à l'IA pourrait avoir des conséquences plus radicales si elle était appliquée à un éventail plus large de domaines, y compris dans des domaines nécessitant des niveaux élevés d'intelligence et de réflexion comme les sciences et l'ingénierie. Compte tenu de la vitesse des ordinateurs qui peut s'avérer nettement plus élevée que celle du cerveau humain (avec des milliards d'opérations par seconde pour une unité de calcul donnée, contre quelques centaines chez l'humain) ainsi que de la possibilité d'étendre les systèmes d'IA à un nombre élevé de matériels informatiques, l'IA générale pourrait donner lieu à de rapides percées dans les domaines scientifique et technique. Certaines de ces percées sont similaires à celles que pourrait réaliser l'être humain, moyennant suffisamment de temps, mais pourraient être accélérées grâce à une IA qui se voit confier la résolution du problème en question. D'autres, par contre, pourraient être impossibles sans l'aide de l'IA en raison des limites cognitives de l'être humain (telles que les limites imposées par la mémoire à long terme et à court terme). Les seules limites précises aux réalisations d'une IA plus sophistiquée sont les limites de la physique, lesquelles permettent le développement d'ordinateurs plus rapides, de matériaux plus solides et la production d'énergie moins coûteuse, notamment par la mise au point de procédés de fabrication de précision atomique (Drexler, 2013). Dans le domaine de la recherche biologique, même le vieillissement ne constitue pas clairement une caractéristique inhérente à la condition humaine et de nombreuses autres améliorations physiques et cognitives semblent physiquement possibles (Kurzweil, 2005).

Amélioration de la coordination

Des systèmes d'IA plus sophistiqués, s'ils sont bien appliqués, pourraient permettre de résoudre certains des conflits sociaux les plus tenaces à l'heure actuelle au moyen d'une coordination améliorée. La société connaît tant de dilemmes du prisonnier et d'autres problèmes liés à l'action collective, dans lesquels le bien-être général de deux parties ou plus se trouverait amélioré par leur coopération mais où ces parties ont des raisons de ne pas coopérer. Par le passé, ces dilemmes ont justifié la mise en place de structures étatiques puissantes et d'organisations internationales chargées de coordonner les gouvernements. Nos outils de coordination sont toutefois limités, d'une part parce qu'il est difficile de surveiller le comportement d'êtres humains pour déceler les signes de violation d'un accord, et d'autre part parce qu'il peut s'avérer difficile d'établir la confiance entre les personnes et les groupes lorsque les intentions des parties sont illisibles. Chacun de ces obstacles à la coopération (surveillance insuffisante et manque de fiabilité des êtres humains) pourrait être atténué par l'application de l'IA dans le cadre de l'exécution des accords. J'aborde chacun d'entre eux tour à tour.

En ce qui concerne la surveillance insuffisante, on observe depuis plusieurs décennies une tendance à recueillir et à analyser davantage de données sur le comportement humain. On observe un recours croissant de l'être humain à l'environnement numérique pour effectuer ses transactions et interactions sociales, ce qui rend leur comportement plus contrôlable pour les entreprises et les pouvoirs publics, que ce soit en bien ou en mal. De même, la présence accrue de caméras (caméras de surveillance spécialisées ou caméras intégrées aux téléphones intelligents et à d'autres appareils) peut permettre un suivi des activités physiques des individus. Les cas d'abus des autorités publiques de surveillance sont bien connus, et la présente analyse ne prétend en rien minimiser leur importance. Cependant, la

surveillance au moyen de systèmes d'IA présente un avantage potentiel considérable: ces systèmes peuvent être utilisés afin de surveiller plus efficacement le respect des accords intra- et internationaux et, potentiellement, de mieux suivre la coopération dans des domaines tels que le contrôle des armements, les mesures de protection de l'environnement et la cybercriminalité. La prolifération nucléaire, par exemple, est un problème persistant, comme en témoignent les conflits internationaux récents portant sur les programmes nucléaires de l'Iran et de Corée du Nord. Une partie du problème est lié au respect des accords internationaux (même d'accords largement bénéfiques tels que ceux relatifs à la non-prolifération) et tient au fait que les activités en ligne et hors ligne, même si elles sont plus faciles à détecter qu'elles ne l'ont jamais été, font néanmoins l'objet d'une surveillance faillible qui laisse le champ libre à des activités illégales telles que la vente clandestine d'informations nucléaires. L'IA pourrait permettre l'avènement d'accords plus efficaces et généralisés en automatisant le processus de collecte et d'analyse d'informations provenant de différentes sourceset, partant, d'opérer une surveillance à une échelle nettement plus grande. À cette fin, l'IA et la robotique pourraient être associées afin, par exemple, d'utiliser des drones de petite taille et peu coûteux afin d'accroître la portée des activités de surveillance.

Deuxièmement, l'IA est capable de libérer les régimes de surveillance et, plus généralement, la gouvernance de certains aspects de partialité et de corruption propres à l'homme, et ce justement car elle peut éliminer le facteur humain de certains processus décisionnels. Contrairement à une personne travaillant par exemple à l'agence de sécurité nationale qui pourrait être tentée d'abuser de ses prérogatives pour des raisons personnelles, le code d'un système d'IA utilisé à des fins de surveillance peut faire l'objet de vérifications afin qu'aucun humain ne consulte des données qu'il n'est pas autorisé à voir, ou qu'aucun humain n'ait accès à des données de surveillance. Dans un scénario allant plus loin encore, le cryptage homomorphe pourrait permettre d'analyser des données cryptées en garantissant que même l'IA ne pourra pas voir les données non cryptées (Trask, 2017). Ces mesures permettraient de négocier et de faire appliquer un plus large éventail d'accords, ce qui pourrait contribuer à éliminer de nombreuses formes de criminalité et à accroître le champ d'action potentiel d'institutions politiques efficaces.

Société des loisirs

Le troisième et dernier avantage essentiel d'une IA avancée abordé dans le présent article est la possibilité de faire naître une société des loisirs prospère et éthique. On ne compte plus les prévisions relatives à la chronologie et à l'ordre de l'automatisation de certains emplois par l'IA, par la robotique et par d'autres technologies (Brundage, 2015; Brynjolfsson et McAfee, 2014; Grace *et al.*, 2017). Je ne prendrai pas position sur le temps qu'il faudra pour que la technologie permette d'automatiser tous les emplois humains, mais j'affirme uniquement que ce scénario est possible en principe et probable à l'avenir. Cette conclusion découle logiquement du fait que la cognition et le comportement humains sont des processus physiques qui pourront tôt ou tard être reproduits par d'autres systèmes physiques, à savoir les ordinateurs et (pour les emplois nécessitant une activité physique) les robots. Si ce niveau de capacité technique devait être atteint, il conviendrait de renégocier d'une façon ou d'une autre le contrat social, ce qui pourrait revêtir différentes formes. Un revenu minimal pourrait être versé à tous les membres de la société afin de garantir un niveau de vie de base. Les citoyens et les gouvernements pourraient aussi s'accorder sur la valeur du travail et décider de maintenir l'une ou l'autre forme d'emploi rémunéré (même superflu technologiquement parlant), par exemple en interdisant l'automatisation de certains emplois. Certains emplois pourraient encore être maintenus dans les cas où le client accorde une valeur intrinsèque au fait que la tâche concernée soit réalisée par un humain plutôt que par un système d'IA. Un scénario possible, que je ne défends pas comme étant le meilleur ni le plus probable mais que je présente uniquement comme un scénario offrant une valeur potentielle élevée, est celui d'une société des loisirs rendue possible par l'IA. Dans une telle société, les individus se concentreraient sur les activités qu'ils jugent intrinsèquement enrichissantes (comme l'art et la création, l'apprentissage, les jeux, l'éducation des enfants ou le temps passé avec des amis ou avec son partenaire

affectif) et ne sont plus tenus de travailler pour préserver un niveau de vie élevé. Le niveau de vie minimal d'une telle société pourrait être nettement plus élevé qu'à l'heure actuelle étant donné que la croissance économique rapide résulterait de l'automatisation complète et que de nombreuses autres limites physiques telles que l'amélioration des capacités cognitives et une nette réduction du coût de production de l'énergie et des biens pourraient être rapidement atteintes.

À quel point cette société des loisirs serait-elle meilleure que nos sociétés actuelles ou que les sociétés qui les ont précédées? Bien meilleure: il est difficile d'estimer le niveau de prospérité qu'une telle société pourrait atteindre. On peut toutefois au moins raisonnablement s'attendre à ce qu'elle soit au moins égale à toutes les sociétés créées jusqu'ici par l'homme étant donné l'absence de limites physiques claires à la capacité de générer ce niveau de vie une fois toutes les tâches automatisées. Dans les cas où une qualité de vie si élevée ne peut simplement se réduire à la production de biens physiques bons marchés, comme c'est probable, il serait également possible d'utiliser une réalité virtuelle immersive et des IA socialement interactives (physiques ou virtuelles) en vue d'un ensemble presque illimité d'expériences. La simple reproduction des niveaux de vie actuels sous forme physique ou virtuelle, et à grande échelle, est encore loin des limites de l'épanouissement potentiel, mais cette réflexion illustre le minimum auquel nous pouvons nous attendre.

Il convient d'évoquer enfin, eu égard à la création d'une société des loisirs éthique, le bien-être des systèmes d'IA eux-mêmes, pour autant que ce concept leur soit applicable. Cette préoccupation mérite qu'on s'y attarde sérieusement, et il faut espérer que les progrès à venir en matière de compréhension de l'intelligence et de la conscience nous aideront à mieux comprendre le paysage des esprits possibles. Une perspective éthique convaincante est que le substrat (à savoir un cerveau ou une puce informatique) ne devrait pas *en soi* servir de fondement à une discrimination entre les êtres humains et les IA (Bostrum et Yudkowski, 2011), même s'il est probable que nous découvrions un jour que le type de substrat influe sur le type de conscience possible. Il est possible d'éviter certains dilemmes éthiques par une conception réfléchie et responsable des systèmes: nous pourrions nous efforcer de concevoir par défaut les systèmes de façon à ce qu'ils soient incapables de ressentir de la souffrance, même s'ils sont conscients (Bryson, 2016). Ces problèmes devront être résolus à long terme, mais une chose est claire: compte tenu de ce que nous savons aujourd'hui, une société des loisirs rendue possible par l'IA semble offrir la *possibilité* de parvenir de manière éthique à un niveau élevé de loisirs et de prospérité. Par contre, d'autres voies menant à une société des loisirs (comme celles créées autrefois par l'exploitation d'esclaves) sont clairement contraires à l'éthique. En outre, en l'absence de capacités technologiques associées à une IA avancée, un niveau de vie inférieur est le meilleur que l'on puisse espérer d'une société des loisirs créée par des moyens politiques. On notera qu'il est envisageable que des degrés de bien-être encore plus élevés soient atteints par les systèmes conçus eux-mêmes, plutôt que par les humains exploitant des systèmes conçus, s'il s'avère que leurs substrats permettent des expériences conscientes de ce type, mais l'univers est suffisamment vaste pour que cette évolution (en principe du moins) ne s'oppose pas à ce que les humains atteignent eux aussi un niveau de vie élevé.

Conclusion: une IA évolutive pour améliorer la prospérité et la civilisation humaine

On peut s'attendre à ce que, tôt ou tard, l'homme invente des systèmes d'IA capables de réaliser toute tâche accomplie par les humains, et bien d'autres encore. Nous ne savons pas combien de temps il faudra, mais les experts s'accordent à dire que c'est possible et beaucoup sont d'avis que ce scénario est probable avant la fin de ce siècle. Que pouvons-nous dire, et que ne pouvons-nous pas dire à propos d'un monde doté de tels systèmes?

Nous ne pouvons affirmer avec certitude que les humains survivront pour en profiter. En effet, même sans IA plus avancée, l'être humain possède (au moins depuis la mise au point d'armes nucléaires) la capacité de se détruire lui-même, et certains avancent des arguments convaincants laissant à penser que l'IA pourrait constituer une autre technologie tout aussi dangereuse (Bostrom, 2014). Il n'est toutefois pas non plus établi que nous ne survivrons pas pour en profiter. Il n'y a aucune contradiction

intrinsèque dans l'existence d'un système artificiel hautement intelligent s'efforçant d'améliorer le bien-être humain sans s'opposer à cette position de subordination ni s'en indigner, et de nombreux chercheurs travaillent activement à faire en sorte que nous finissions en définitive par concevoir des systèmes de ce type. Il n'est pas non plus possible d'affirmer que, si nous survivons assez longtemps pour connaître ce monde, il profitera aux humains. Cette technologie pourrait être détournée de manière à créer un État autoritaire stable d'une longévité sans précédent à l'échelle mondiale, fondé sur l'automatisation de la surveillance, la coercition et la répression des opinion divergentes. Par ailleurs, entre l'utopie et la dystopie, de nombreux autres scénarios sont possibles.

Nous pouvons toutefois affirmer plusieurs choses sur les sociétés que l'humanité *pourrait* créer si elle parvient à surmonter cette transition. Les trois facteurs évoqués ci-dessus (accélération des tâches, amélioration de la coordination et société des loisirs) revêtent chacun une importance propre mais, ensemble, ils esquissent une voie vers une civilisation vaste, prospère et présente dans l'espace. Dans un monde où n'importe quelle tâche peut être accélérée à l'aide de l'IA, une tâche dont l'accélération serait largement bénéfique est le développement et le déploiement de technologies permettant la colonisation rapide de l'espace. Cette colonisation donnerait un nouvel accès à un nombre considérable de terres et de ressources matérielles et offrirait à l'humanité des possibilités passionnantes d'exploration. L'ouverture de ces nouvelles frontières associée à la réalisation d'autres tâches, comme la mise au point de nouvelles techniques d'amélioration de nos capacités cognitives et la production de biens et de services nettement moins chers permettrait une nouvelle Renaissance de l'humanité. Même si l'IA pourrait devenir (ou servir à créer) une nouvelle génération d'armes dressant les États et les particuliers les uns contre les autres, elle pourrait aussi servir à négocier des accords internationaux (voire, à terme, interplanétaires) ambitieux visant à interdire ces utilisations malveillantes.

Cette nouvelle Renaissance pourrait échouer pour de nombreuses raisons. Nous pourrions nous quereller sur les gains relatifs que nous tirerions de l'IA et nous embourber dans des conflits internationaux en perdant de vue les gains absolus nettement plus importants et accessibles à tous, mais nous pourrions aussi mettre en place un système d'IA qui semble de prime abord traduire nos valeurs et qui, au final, entraîne une stagnation culturelle et affaiblisse l'humanité. Toutefois, comme dans le cas des difficultés techniques évoquées ci-dessus, je ne vois aucune raison de penser que ces défis politiques sont insurmontables.

Remerciements

Je remercie John Danaher, Anders Sandberg, Ben Garfinkel, Carrick Flynn, Stuart Armstrong et Eric Drexler pour leurs commentaires précieux sur des versions antérieures de ces notes. Toute erreur résiduelle relève de la responsabilité de l'auteur.

Références

AI Impacts, 2017. «AI hopes and fears in numbers», blog *AI Impacts*, https://aiimpacts.org/ai-hopes-and-fears-in-numbers/

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., et Mané, D. 2016. «Concrete Problems in AI Safety», arXiv preprint server, https://arxiv.org/abs/1606.06565

Armstrong, S., Bostrom, N., and Shulman, C. 2016. «Racing to the Precipice: a Model of Artificial Intelligence Development», *AI & Society*, pp. 1-6.

Bostrom, N. alors Yudkowsky, E. 2011. «The Ethics of Artificial Intelligence», in *Cambridge Handbook of Artificial Intelligence*, ed. Ramsey, W. et Frankish, K.

Bostrom, N. 2014. *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press.

Bostrom, N. 2017. «Strategic Implications of Openness in AI Development», *Global Policy*, Vol. 8, Numéro 2.

Bostrom, N., Dafoe, A., et Flynn, C. 2017. «Policy Desiderata in the Development of Machine Superintelligence», http://www.nickbostrom.com/papers/aipolicy.pdf

Brundage, M. 2016. «Economic Possibilities for Our Children: Artificial Intelligence and the Future of Work, Education, and Leisure», *2015 AAAI Workshop on AI, Ethics, and Society*.

Brundage, M. et Avin, S. *et al.* 2018. «The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation.»

Bryson, J. 2016. «Patiency is Not a Virtue: AI and the Design of Ethical Systems», *2016 AAAI Spring Symposium Series*.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., et Evans, O. 2017. «When Will AI Exceed Human Performance? Evidence from AI Experts», arXiv preprint server, https://arxiv.org/abs/1705.08807

Kirkpatrick, K. 2016. «Battling Algorithmic Bias», *Communications of the ACM*, Vol. 59, No. 10, pp. 16-17, https://cacm.acm.org/magazines/2016/10/207759-battling-algorithmic-bias/abstract

Kurzweil, R. 2005. *The Singularity is Near: When Humans Transcend Biology*. New York: Viking Press.

Trask, A. 2017. «Safe Crime Detection», https://iamtrask.github.io/2017/06/05/homomorphic-surveillance/

5. Observations sur l'intelligence artificielle et l'optimisme rationnel

Olle Häggström

Introduction

L'avenir de l'intelligence artificielle (IA) et son incidence sur l'humanité sont des sujets importants. Ils ont été abordés lors d'un débat organisé à Bruxelles par le panel de l'évaluation des choix scientifiques et technologiques (STOA) du Parlement européen le 19 octobre 2017. Steven Pinker était le principal intervenant de cette manifestations, accompagné de Peter Bentley, Miles Brundage, Thomas Metzinger et moi-même (voir vidéo du panel de la STOA, 2017). Le présent essai s'inspire de mes travaux préparatoires à cette manifestation et de réflexions tirées en partie d'un article de mon blog (Häggström, 2017) consacré aux interventions des autres membres du panel à cette occasion.

Optimisme

L'intitulé de la manifestation organisée le 19 octobre reprenait l'expression «optimisme rationnel», qui m'a semblé contradictoire puisque je considérais l'optimisme et le pessimisme comme des distorsions biaisées de données factuelles disponibles. Il me paraissait en particulier *irrationnel* d'affirmer, en s'appuyant sur des preuves insuffisantes, que tout allait bien se passer pour l'humanité. À mieux y réfléchir pourtant, j'ai décidé qu'il existait un autre type d'optimisme que je suis plus enclin à qualifier de rationnel, à savoir ...

> ... le fait d'avoir une vision épistémique bien balisée de l'avenir et de ses incertitudes, d'accepter le fait que l'avenir n'est pas gravé dans le marbre et d'agir en se fondant sur l'hypothèse de travail selon laquelle la probabilité d'un avenir heureux peut dépendre des actes que nous posons aujourd'hui.

Il convient de noter que cette hypothèse de travail pourrait se révéler (du moins partiellement) incorrecte. Le monde pourrait être tellement chaotique qu'il serait vain d'essayer de déterminer si un acte particulier posé aujourd'hui est susceptible d'augmenter ou de réduire les chances d'un avenir long et prospère pour l'humanité. Si tel est le cas, nos actions n'ont pas d'importance (prévisible) pour cet avenir. Nous ne savons néanmoins pas si tel est le cas, et il est donc logique de *supposer* (sans certitude) que nos actes ont de l'importance et de nous efforcer de déterminer les actions qui améliorent nos chances d'un avenir radieux. C'est dans cet esprit qu'a été écrit le reste de cet essai.

Intelligence artificielle

Comme toute autre technologie émergente, comme la biologie synthétique et les nanotechnologies, l'IA comporte à la fois des avantages notables et des risques considérables. En ce qui concerne les avantages, le cabinet de consultants en gestion McKinsey & Co a publié en 2013 un rapport estimant à 50 000 milliards de dollars la valeur ajoutée économique des innovations en matière d'IA et de robotique au niveau mondial sur les dix prochaines années (Manyika *et al.* 2013; Omohundro, 2015), ce qui constitue à mon sens un chiffre trop prudent qui s'explique en partie par le rythme inattendu auquel l'apprentissage des machines, alimenté par les données massives, a décollé depuis 2013. Nous ne devrions certes pas commettre l'erreur de supposer que la croissance économique est automatiquement synonyme d'amélioration de la vie, mais il est clair que les progrès en matière d'IA peuvent nous apporter de nombreuses choses. À plus long terme, rares sont les limites (hormis les lois de la physique) aux bienfaits que peut procurer l'IA.

Il existe différents types de risques. Le risque le plus étroitement lié aux avantages économiques estimés est l'incidence que peut avoir une automatisation fondée sur l'IA sur le marché du travail. En ce qui concerne les véhicules autonomes, l'intégralité d'un secteur du marché du travail comptant notamment des millions de chauffeurs de poids-lourds, de bus et de taxis risque de disparaître entièrement en 20 ans

à peine. Toutes ces personnes trouveront-elles un emploi ailleurs ou bien se retrouveront-elles au chômage? D'autres secteurs du marché du travail risquent de subir le même sort. Par ailleurs, si le remplacement de travailleurs humains n'est évidemment pas un phénomène nouveau, la révolution de l'IA va plus loin: les machines ne remplacent plus uniquement le travail manuel, mais aussi de plus en plus au travail intellectuel. Associé au rythme accru de l'automatisation, ce phénomène soulève de sérieux doutes quant à la possibilité de trouver de nouvelles tâches pour les travailleurs humains à un rythme permettant de compenser l'automatisation (comme ce fut le plus souvent le cas par le passé) ou si l'on doit craindre une envolée du chômage. Voir par exemple à ce sujet le livre de Brynjolfsson et McAfee (2014). À long terme, le scénario extrême dans lequel les machines sont plus performantes que l'homme dans tous les emplois, avec pour résultat un taux de chômage de 100 %, n'est peut-être pas irréaliste. Au moins deux questions de société essentielles découlent de cette réflexion. Tout d'abord, comment organiser une société dans laquelle les gens ne travaillent pas mais consacrent leur temps à des aspirations plus élevées telles que l'art, la culture, l'amour ou, tout simplement au plaisir fou de jouer à des jeux vidéos? Deuxièmement, même si nous parvenons à concevoir cette utopie, il reste à négocier la transition de la société actuelle à cette utopie sans créer des inégalités économiques et des troubles sociaux sans précédent.

Si cela semble modérément alarmant, imaginez les conséquences de l'évolution de l'IA pour les armements autonomes. Permettez-moi de citer simplement un passage d'une lettre ouverte que j'ai cosignée avec des milliers d'autres scientifiques en 2015 (Russel *et al.* 2015):

> Si une grande puissance militaire se lance dans la mise au point d'armes utilisant l'intelligence artificielle, une course mondiale aux armements est pratiquement inévitable et son issue est inéluctable: les armes autonomes deviendront les kalachnikovs de demain. Contrairement aux armes nucléaires, elles ne nécessitent pas de matières premières coûteuses ou difficiles à obtenir. Elles deviendront donc omniprésentes, et toutes les puissances militaires importantes les produiront en masse à moindres coûts. Elles finiront tôt ou tard sur le marché noir et aux mains de terroristes, de dictateurs désireux de mieux contrôler leurs populations, de seigneurs de la guerre en quête d'épuration ethnique, etc. Les armes autonomes sont idéales pour mener à bien des missions telles que les assassinats, la déstabilisation de nations, la répression de populations et le massacre sélectif de groupes ethniques particuliers. Nous croyons par conséquent qu'une course aux armements dotés de l'IA ne serait pas bénéfique pour l'humanité.

Lors de la réunion à Bruxelles (STOA, 2017, à 12:01:00 selon l'horloge de la vidéo), Steven Pinker a présenté un point de vue plus optimiste quant à ce risque lié aux armements à intelligence artificielle. Il a écarté cette possibilité en affirmant que seul un fou voudrait créer quelque chose d'aussi horrible qu'un «essaim de robots conçus pour attaquer des individus sur la base d'une reconnaissance faciale» et qu'aucun dictateur fou n'aurait plus la possibilité de faire une telle chose à l'heure actuelle, car l'ingénierie n'est plus le fait de génies solitaires mais de grandes collaborations. Cette vision naïve fait fi du mode de fonctionnement des courses aux armement et du complexe militaro-industriel ainsi que du fait que nous développons des armes de destruction massive tout aussi horribles depuis plus de 70 ans. Ces évolutions ne résultent pas du travail de fous solitaires, mais de vastes activités de collaboration (le Projet Manhattan étant le plus célèbre), alors pourquoi penser que cette tendance va soudain disparaître? L'objection de Pinker relève manifestement de l'«optimisme irrationnel» que j'évoquais plus haut.

Ces deux risques (risque d'inégalités économiques dues à la montée du chômage et risque d'une course aux armements dotés de l'IA) doivent être pris au sérieux, et il convient de déterminer la mesure de leur gravité et la manière de les atténuer. Dans les trois prochains chapitres, j'aborde plus en profondeur un troisième type de risque lié à l'IA – plus lointain et hypothétique peut-être que les deux autres, mais non moins réel pour autant: le risque de voir apparaître une IA superintelligente dont les valeurs ne correspondent pas aux nôtres.

Risque engendré par une superintelligence

Imaginons que les chercheurs atteignent un jour leur objectif de longue date, celui de créer une IA superintelligente, à savoir une machine qui dépasse clairement les humains dans tous les domaines de compétences que nous qualifions d'intelligence. À ce stade, nous ne pouvons plus espérer garder le contrôle. L'expérience de pensée connue sous le nom d'*Armageddon du trombone* peut faire office de mise en garde (Bostrom, 2003):

Imaginons une usine à trombones à papier dirigée par une IA avancée (mais pas encore superintelligente) en vue de maximiser la production de trombones. Ses informaticiens s'efforcent en permanence de l'améliorer et, un jour, plus ou moins par hasard, ils parviennent à améliorer la machine à un point tel qu'elle entre dans une spirale incontrôlée d'auto-amélioration connue sous le nom d'*explosion de l'intelligence* ou de *Singularité*. Elle devient rapidement la première IA superintelligente au monde et, après avoir atteint son objectif, à savoir maximiser la production de trombones, elle poursuit sur sa lancée et transforme la planète entière (nous compris) en un immense tas de trombones à papier avant de se lancer à la conquête de l'espace pour transformer le système solaire, la Voie lactée et le reste de l'univers observable en trombones.

Cet exemple se veut caricatural pour illustrer d'autant mieux un phénomène beaucoup plus général (à ma connaissance, personne ne craint réellement qu'une IA ne transforme le monde en trombones). Il vise à mettre en lumière qu'aucune intention malveillante n'est nécessaire pour qu'une percée dans l'IA devienne dangereuse: nous n'avons pas besoin d'une histoire mettant en scène un savant fou qui cherche à détruire la planète pour se venger de l'humanité. Même des ambitions apparemment innocentes, comme maximiser la production de trombones, peuvent aboutir à des scénarios dangereux.

Mais est-ce vraiment le cas? Deux des membres du panel de la réunion de Bruxelles (Pinker et Bentley) ont défendu avec force le point de vue selon lequel le risque d'une catastrophe causée par une superintelligence ne mérite pas d'être pris au sérieux. Ils semblaient ravis d'être d'accord sur ce point, même si les raisons qu'ils invoquaient étaient très différentes.

Pour déterminer si le risque de catastrophe provoquée par une superintelligence est réel, il convient de scinder le problème en deux parties:

> (1) Peut-on s'attendre à ce que l'évolution de l'IA aboutisse un jour à la création d'une superintelligence? Dans l'affirmative, quand et à quel rythme?
>
> (2) Une fois créée, qu'est-ce que cette IA superintelligente serait susceptible de faire? Risque-t-elle de faire quelque chose de dangereux?

J'aborde ces deux questions séparément dans les deux chapitres suivants. Pour que le risque d'une superintelligence soit réel, il faut répondre «oui» à la première question et «oui, elle risquerait de faire quelque chose de dangereux» à la deuxième question. Lors de la réunion qui s'est tenue à Bruxelles, Bentley a mis en doute la réponse à la première question et Pinker la réponse à la deuxième question.

Pour quant peut-on (éventuellement) attendre une superintelligence?

Si l'on adopte une vision naturaliste du monde (à savoir que l'esprit humain ne naît pas par dualisme cartésien, d'une intervention divine ou d'un autre tour de magie), on peut raisonnablement s'attendre à ce que lorsque l'évolution biologique a résulté en la création du cerveau humain, elle était bien loin de la manière généralement optimale de configurer la matière pour en tirer une intelligence maximale. Il faut donc s'attendre à ce que d'autres configurations de matière aboutissent à une superintelligence. Dès lors, on peut aisément conclure (en s'appuyant par exemple sur la thèse de Church) qu'il est possible de simuler cette combinaison sur un ordinateur, auquel cas il est en principe possible de créer une superintelligence grâce à un programme informatique adapté.

À quel point serait-il difficile de définir un tel programme? Nous l'ignorons. Ces dernières années en particulier, les évolutions en matière d'IA ont été très prometteuses notamment en ce qui concerne la création d'IA consacrées à des tâches spécifiques telles que conduire une voiture ou battre des joueurs humains aux échecs ou au jeu de Go. Les progrès accomplis en vue de la création d'une intelligence *générale* artificielle (IGA) – une machine faisant preuve d'une intelligence égale ou supérieure à celle de l'homme de manière suffisamment souple pour fonctionner dans tous les domaines typiquement rencontrés par l'être humain (échecs, basket, développement logiciel, cuisine, soins infirmiers, reconnaissance faciale, conversation ordinaire, etc.) – sont nettement moins impressionnants. Certains affirment que les progrès accomplis sont inexistants, mais ces propos me semblent quelque peu injustes. Par exemple, une IA a été mise au point il y a quelques années et a rapidement appris à jouer à une série de jeux vidéo Atari (Clark, 2015). Même s'il est vrai que nous sommes encore loin de pouvoir gérer toutes les tâches auxquelles les humains se trouvent confrontés dans le monde réel, il s'agit d'une amélioration tangible par rapport à une compétence spécialisée dans un seul jeu. L'une des nombreuses façons possibles d'arriver à une IGA consiste à élargir progressivement le domaine dans lequel une machine est capable d'agir intelligemment.

Il existe de nombreuses approches possibles de la création d'un logiciel intelligent. On observe actuellement un grand bond en avant de l'«apprentissage profond» (LeCun *et al.* 2015), qui consiste essentiellement en une renaissance et un développement plus approfondi d'anciennes techniques de réseaux neuronaux qui donnaient autrefois des résultats peu convaincants mais qui aujourd'hui, grâce à la vitesse des ordinateurs et à la disponibilité de jeux de données énormes pour entraîner les machines permettent de résoudre des problèmes complexes les uns à la suite des autres. Il s'agit d'un exemple de système de «boîte noire», selon lequel, les ingénieurs qui parviennent à créer une IA ne comprennent bien souvent pas son raisonnement. Un autre exemple de système de boîte noire est la programmation génétique, dans le cadre de laquelle un ensemble de logiciels candidats concourent en reproduisent les mécanismes de sélection-reproduction-mutation de l'évolution biologique. Il existe toutefois d'autres approches, différentes du système de boîte noire, notamment la «GOFAI» («Good Old-Fashioned AI», bonne vieille IA) pour laquelle les programmeurs encodent manuellement eux-mêmes les concepts et procédures de raisonnement de la machine. Il peut également exister des méthodes qui reposent sur l'imitation du cerveau humain, que ce soit par la compréhension du type de traitement de haut niveau des informations par le cerveau qui constitue la clé de l'IGA, ou [approche défendue avec force par Kurzweil (2005)] par l'utilisation de la force brute pour copier le fonctionnement exact du cerveau à un niveau de détail suffisant (au niveau des synapses, voire à un niveau inférieur encore) pour en reproduire le comportement.

Il est possible qu'aucune de ces approches n'aboutisse jamais à une IGA, mais il semble raisonnable d'envisager au moins la possibilité que l'une d'entre elles, ou qu'une combinaison d'approches, aboutisse finalement à une IGA. Mais quand? Cela semble plus incertain encore, et une enquête réalisée par Müller et Bostrom (2016) recueillant les estimations des 100 chercheurs en IA les plus cités au monde a réparti les estimations sur l'intégralité de ce siècle (et au-delà). Leur estimation médiane pour l'émergence de ce que l'on pourrait appeler une IGA de niveau humain est 2050, avec une estimation médiane pour l'émergence d'une superintelligence dans les 30 ans qui suivent. Voir également l'enquête plus récente (Grace *et al*., 2017). Compte tenu des avis très divergents des experts, il serait téméraire du point de vue épistémique de défendre une conviction ferme quant à la date d'émergence d'une superintelligence. Il convient plutôt d'admettre avec prudence et réflexion qu'elle pourrait apparaître au cours des prochaines décennies, des prochains siècles, voire jamais.

Et pourtant, lors de la réunion de Bruxelles, Peter Bentley s'est exprimé à son sujet et a déclaré que «ça ne se produira jamais, c'est tout! C'est complètement irrationnel de ne fût-ce qu'imaginer qu'elle puisse apparaître» (STOA, 2017, à 12:08:45). D'où provient donc cette certitude? Au cours de sa présentation, Bentley n'a avancé qu'un seul argument à l'appui de sa position, à savoir son expérience et celle d'autres développeurs en IA selon laquelle tout progrès dans ce domaine nécessite de travailler dur et tous les

nouveaux algorithmes inventés ne permettent de résoudre qu'un problème particulier. Une fois l'objectif atteint, l'amélioration d'abord rapide de l'algorithme serait toujours suivie d'une période de rendements décroissants. Voilà pourquoi, a-t-il souligné, la résolution d'un autre problème nécessite toujours de travailler dur pour inventer et appliquer un autre algorithme.

L'argumentation qu'avance Bentley ignore un fait bien connu, à savoir qu'il existe des algorithmes présentant une capacité de résolution de problèmes moins limitée, comme en témoigne le logiciel du cerveau humain. Sa conviction absolue selon laquelle l'ingéniosité scientifique humaine ne parviendra pas à découvrir un tel algorithme au cours du siècle à venir (ou même sur n'importe quelle échelle de temps) semble difficile à défendre du point de vue rationnel: elle demande une foi dogmatique.

Pour synthétiser le présent chapitre: Même s'il subsiste une possibilité que l'IA n'aboutisse jamais à une superintelligence, il est aussi tout à fait plausible qu'elle y parvienne. En supposant qu'elle y parvienne, il est impossible de prédire quand et, pour tenir compte de cette incertitude, nous devons reconnaître que la superintelligence pourrait naître à n'importe quel moment au cours ce de siècle ou peut-être même plus tard. Comme le souligne un article important rédigé par Sotala et Yampolskiy (2015), nous devons éviter de tomber dans le piège et de penser que, parce que la date d'apparition de la superintelligence est incertaine, elle doit également être lointaine.

Qu'est-ce qu'une IA superintelligence décidera de faire?

Imaginons donc une situation future dans laquelle une IA superintelligente a été créée, scénario qui, comme je l'ai indiqué au chapitre précédent, est tout à fait plausible. Il semble fort probable que, dans une telle situation, nous perdrons le contrôle et que notre destin dépendra de ce que cette IA décidera de faire, de même qu'aujourd'hui, le destin des chimpanzés dépend des décisions prises par les humains et non plus de leurs propres décisions. Pour éviter d'arriver à une telle situation, il convient, entre autres, d'encadrer l'IA et de l'empêcher d'influencer le monde d'une autre manière qu'au moyen d'un canal de communication étroit soigneusement contrôlé par des administrateurs humains de la sécurité. Cette approche de «l'IA enfermée dans une boîte» a suscité quelque peu d'intérêt dans le secteur de la recherche en matière de sécurité de l'IA (voir par ex. Armstrong *et al.*, 2012), mais la conclusion générale est souvent que garder le contrôle d'une IA superintelligente serait trop difficile pour de simples humains et que nous pouvons au mieux espérer contrôler cette IA de manière provisoire et pour une courte période.

Imaginons donc également que cette IA superintelligente est sortie de sa boîte de pandore et qu'elle est en mesure de circuler librement sur l'internet (y compris l'internet des objets). Elle peut se copier en de multiples exemplaires, utiliser son intelligence supérieure pour traverser tous les pare-feux mis en travers de son chemin, etc. Nous avons alors perdu le contrôle, et la survie et le bien-être futurs de l'humanité dépendent de ce que la machine décide de faire. Que va-t-elle donc décider? Tout dépendra de ses objectifs. C'est difficilement prévisible et tout débat sur ce point est forcément en partie spéculatif. Il existe toutefois un cadre nous permettant d'aller au-delà de la simple spéculation, à savoir ce que j'ai décidé d'appeler (Häggström, 2016) la *théorie Omohundro-Bostrom des objectifs ultimes et instrumentaux de l'IA* (Omohundro, 2008; Bostrom, 2012, 2014). Cette théorie n'est pas gravée dans le marbre comme le serait un théorème mathématique reconnu et est donc susceptible de révision, tout comme les prévisions qu'elle permettra éventuellement de formuler. Elle est toutefois suffisamment plausible pour que ses prévisions méritent d'être prises au sérieux. Elle repose sur deux éléments essentiels: la thèse d'orthogonalité et la thèse de convergence instrumentale. Permettez-moi de les expliquer tour à tour.

La *thèse d'orthogonalité* affirme (en gros) qu'à peu près n'importe quel objectif ultime est compatible avec des degrés arbitrairement élevés d'intelligence. Il est possible de constituer des contre-exemples qui s'inspireraient de l'idée de paradoxes autoréférentiels (par exemple «gardez votre niveau d'intelligence général inférieur à celui d'un chien moyen en 2017»), mais l'idée est qu'il est également possible de programmer n'importe quelle fonction cible à optimiser par votre IA, et cet objectif peut être attribué à des IA de n'importe quel degré d'intelligence. Ceux pour qui la théorie Omohundro-Bostrom et la

futurologie en matière d'IA de manière générale sont des concepts nouveaux avanceront souvent l'objection qu'un objectif apparemment étroit tel que maximiser la production de trombones est intrinsèquement stupide et qu'il est donc contradictoire de suggérer qu'une IA superintelligente pourrait avoir un tel objectif. Mais cela revient à confondre intelligence et objectifs: l'intelligence est uniquement la capacité à orienter le monde vers des objectifs spécifiques, quels qu'ils soient. Maximiser la production de trombones nous *semble* stupide, non pas parce cet objectif est objectivement stupide, mais parce qu'il est contraire à *nos* propres objectifs.

Vient ensuite la *thèse de la convergence instrumentale.* L'IA peut adopter différents objectifs instrumentaux, non pas pour eux-mêmes, mais en tant qu'outils permettant de réaliser son objectif ultime. La thèse de la convergence instrumentale affirme qu'il existe un certain nombre d'objectifs instrumentaux que l'IA devrait logiquement adopter en vue de réaliser les objectifs extrêmement variés qu'elle pourrait avoir. Les principaux objectifs auxquels cette thèse semble s'appliquer sont notamment:

- l'auto-préservation (ne les laissez pas vous mettre hors fonction!);
- l'acquisition de matériel et d'autres ressources;
- l'amélioration de ses propres logiciel et matériel;
- la préservation de l'objectif ultime; et
- si l'objectif ultime n'est pas en phase avec les valeurs humaines, faire profil bas (cacher ses objectifs et/ou ses capacités) jusqu'au moment où il est possible de surmonter aisément toute résistance humaine.

Un cas typique du fonctionnement de cette logique est le premier objectif principal de la liste: l'auto-préservation. Quel que soit son objectif ultime, l'IA considèrera probablement qu'elle a plus de chances d'atteindre cet objectif si elle existe et si elle fonctionne, plutôt que si elle est détruite ou désactivée. Il est donc logique que l'IA résiste à nos tentatives de la désactiver. Un raisonnement similaire permet de démontrer les autres objectifs principaux de la liste. L'objectif principal d'amélioration de ses propres logiciel et matériel est ce qui devrait logiquement pousser l'IA, une fois qu'elle a acquis l'intelligence nécessaire pour concevoir des IA, à entrer dans la spirale d'auto-amélioration mentionnée ci-avant. Cette spirale peut ou ne peut pas être suffisamment rapide (selon que les rendements du réinvestissement cognitif augmentent ou diminuent; voir Yudkowsky, 2013) pour mériter la qualification d'«explosion de l'intelligence».

Le concept de convergence de moyens échappe bien souvent aux critiques du discours relatif au risque lié à la superintelligence. Lors de la réunion de Bruxelles en particulier, j'ai été déçu d'entendre Pinker affirmer ce qui suit, quelques minutes à peine après avoir expliqué les bases de la théorie Omohundro-Bostrom et le cas particulier de l'auto-préservation:

> Si nous donnions [à la machine] l'objectif de se préserver elle-même, elle ferait n'importe quoi pour se préserver, y compris nous détruire. [...] Pour l'éviter, une seule chose à faire: ne pas créer de systèmes si stupides! (STOA, 2017, 11:57:45)

Ce point de vue passe à côté de l'essence, à savoir qu'en vertu de la théorie Omohundro-Bostrom, une IA suffisamment intelligente adoptera probablement l'objectif principal d'auto-préservation, indépendamment du fait que cet objectif lui ait ou non été attribué explicitement par ses programmeurs humains.

Le cas de la préservation de l'objectif ultime est particulièrement intéressant. On pourrait être tenté de croire qu'une IA chargée de maximiser la production de trombones, une fois qu'elle atteindra un niveau d'intelligence suffisant, percevra l'étroitesse et la stupidité de cet objectif et passera à autre chose. Imaginons donc que cette IA envisage d'adopter un autre objectif plus louable (à nos yeux) tel que la préservation des écosystèmes. Elle se demandera «qu'est-ce qui est préférable, m'en tenir à la maximisation des trombones ou m'atteler à la préservation des écosystèmes?». Mais qu'entend-on par «préférable» dans ce contexte? Quel est le critère permettant de déterminer lequel de ces objectifs est préférable? Eh bien, puisque l'IA n'a pas encore changé d'objectif mais en est encore à envisager de le

faire, son objectif reste de maximiser la production de trombones. Son critère d'évaluation sera donc «quel objectif permettra de produire le plus grand nombre de trombones?». La réponse à cette question est très probablement «la maximisation des trombones», de sorte que l'IA s'en tiendra à cet objectif. Il s'agit là du mécanisme de base qui sous-tend l'objectif principal de préservation de l'objectif ultime.

Du fait de ce mécanisme, il est peu probable qu'une IA superintelligence nous permette de manipuler son objectif ultime. Par conséquent, si cet objectif ultime est de maximiser la production de trombones, nous sommes probablement condamnés. Nous devons donc attribuer à l'IA les objectifs qui nous semblent préférables avant qu'elle n'atteigne la superintelligence. Tel est l'objectif du programme de recherche sur l'*alignement de l'IA*, formulé (sous l'appellation alternative l'*IA amicale*, qu'il est sans doute préférable d'éviter vu ses connotations inutilement anthropomorphiques) dans un article fondateur de 2008 rédigé par Yudkowsky (2008) et qui a suscité de nombreux débats depuis lors (voir par ex. Bostrom, 2014; Häggström, 2016; Tegmark, 2017) . Pour s'attaquer de manière systématique à ce problème, il est possible de le scinder en deux: premièrement, l'aspect technique d'assigner les objectifs souhaités à l'IA; deuxièmement, l'aspect éthique de définir ces objectifs et/ou de savoir qui a le droit de les définir et au moyen de quelle procédure (démocratique ou autre). Ces deux problèmes sont extrêmement difficiles. Par exemple, nous savons au moins depuis Yudkowsky (2008) que les valeurs humaines sont très fragiles dans la mesure où la moindre petite erreur peut entraîner des catastrophes aux mains d'une IA superintelligente. La raison pour laquelle nous devons travailler à l'alignement de l'IA dès aujourd'hui n'est pas que la superintelligence est sur le point d'apparaître (malgré Yudkowski, 2017) mais plutôt que, même si elle n'apparaît que dans plusieurs décennies, nous pourrions bien avoir besoin de ces décennies sans guère de temps à perdre.

Lorsque Pinker, dans le passage cité ci-dessus, affirme que «[p]our l'éviter, une seule chose à faire: ne pas créer de systèmes si stupides!», on peut l'interpréter comme un plaidoyer *en faveur* de l'alignement de l'IA.s Je pense toutefois que cette formulation ne traduit pas la difficulté du problème et donne l'impression erronée que l'alignement de l'IA ne nécessite pas une attention sérieuse.

Devrions-nous éviter d'en parler?

Dans son argumentation de la réunion de Bruxelles visant à ignorer le risque d'une IA apocalyptique, Pinker a fait remarquer (STOA, 2017, 11:51:40) que le grand public doit déjà s'inquiéter de la menace nucléaire et du changement climatique. Il affirme que, de ce fait, un risque mondial supplémentaire est susceptible de nous abattre et de nous amener à perdre tout espoir pour l'avenir. Cette spéculation n'est peut-être pas sans fondement, mais pour évaluer la pertinence de cet argument, nous devons examiner séparément les deux possibilités: (a) un risque réel d'IA apocalyptique et (b) l'inexistence du risque d'IA apocalyptique.

Si la possibilité (b) est retenue, il est *évident* que nous ne devons pas gaspiller notre temps et notre énergie à discuter de ce risque, mais cela n'a rien à voir avec une quelconque nécessité d'épargner la sensibilité du public. Imaginons plutôt que la possibilité (a) soit d'application. Dans ce cas, la recommandation de Pinker revient à ignorer une menace susceptible de tous nous tuer, ce qui ne me semble pas une bonne idée. Il serait évidemment formidable de survivre à la menace nucléaire et de résoudre la crise climatique, mais cela ne nous aidera guère si nous allons droit vers une apocalypse provoquée par l'IA. Le fait de passer sous silence un risque réel semble également contraire à l'une des idées les plus importantes aux yeux de Pinker depuis au moins dix ans, à savoir celle de l'ouverture scientifique et intellectuelle, et des valeurs du siècle des Lumières de manière plus générale. Il en va de même en cas d'impossibilité de déterminer avec certitude si (a) ou (b) est d'application. Dans ce cas, l'approche la plus conforme aux valeurs du siècle des Lumières est de discuter ouvertement du problème et de tenter de déterminer si le risque est réel.

Conclusions et lectures supplémentaires

Si nous nous y préparons avec le soin nécessaire, l'apparition d'une superintelligence pourrait être la meilleure chose qui ne soit jamais arrivée à l'humanité, mais elle s'accompagne également d'un risque élevé de catastrophe. Ce risque ainsi que les risques plus prosaïques liés à l'IA évoqués précédemment méritent notre attention. L'apocalypse provoquée par l'IA n'est pas une certitude, mais il s'agit d'une possibilité suffisamment plausible pour que nous nous efforcions de trouver comment l'éviter. Tel est mon argument dans cet essai. J'ai cependant été fort bref, et je conseille au lecteur en quête d'un développement plus poussé de cet argument de consulter le chapitre 4 de mon livre (Häggström, 2016). Pour une analyse plus détaillée encore, je recommande vivement les livres de Bostrom (2014) et de Tegmark (2017). Le livre de Tegmark s'adresse plus clairement à un large public, tandis que celui de Bostrom est plus académique, mais tous deux regorgent d'idées étonnantes et importantes (dont certaines reviennent dans les deux livres).

Remerciement

Je remercie Björn Bengtsson pour ses observations précieuses sur le manuscrit.

Références

Armstrong, S., Sandberg, A. et Bostrom, N. (2012) Thinking inside the box: controlling and using an oracle AI, *Minds and Machines* 22, 299-324.

Bostrom, N. (2003) Ethical issues in advanced artificial intelligence, *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence, Vol. 2* (ed. Smit, I. et al.) International Institute of Advanced Studies in Systems Research and Cybernetics, pp. 12-17.

Bostrom, N. (2012) The superintelligent will: motivation and instrumental rationality in advanced artificial agents, *Minds and Machines* 22, 71-85.

Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, Oxford.

Brynjolfsson, E. et McAfee, A. (2014) *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, W.W. Norton, New York.

Clark, L. (2015) DeepMind's AI is an Atari gaming pro now, *Wired*, 25 février.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B. et Evans, O. (2017) When will AI exceed human performance? Evidence from AI experts, *arXiv*:1705.08807.

Häggström, O. (2016) *Here Be Dragons: Science, Technology and the Future of Humanity*, Oxford University Press, Oxford.

Häggström, O. (2017) The AI meeting in Brussels last week, *Häggström hävdar*, 23 octobre.

Kurzweil, R. (2005) *The Singularity Is Near: When Humans Transcend Biology*, Viking, New York.

LeCun, Y., Bengio, Y. et Hinton, G. (2015) Deep learning, *Nature* 521, 436-444.

Manyika, J., Chui, M., Bughin, J., Dobbs, R. Bisson, P. et Marrs, A. (2013) Disruptive technologies: Advances that will transform life, business, and the global economy, *McKinsey Global Institute*.

Müller, V. & Bostrom, N. (2016) Future progress in artificial intelligence: A survey of expert opinion. In *Fundamental Issues of Artificial Intelligence,* Springer, Berlin, pp. 553-571.

Omohundro, S. (2008) The basic AI drives, *Artificial General Intelligence 2008: Proceedings of the First AGI Conference* (Wang, P., Goertzel, B. et Franklin, S., eds), IOS, Amsterdam, pp 483-492.

Omohundro, S. (2015) McKinsey: $50 trillion of value to be created by AI and robotics through 2025, *Self-Aware Systems*, 4 août.

Russell, S. et al. (2015) *Autonomous Weapons: An Open Letter from AI and Robotics Researchers*, Future of Life Institute.

Sotala, K. et Yampolskiy, R. (2015) Responses to catastrophic AGI risk: a survey, *Physica Scripta* 90, 018001.

STOA (2017), vidéo de la réunion STOA du 19 octobre 2017, https://web.ep.streamovations.be/index.php/event/stream/171019-1000-committee-stoa/embed

Tegmark, M. (2017) *Life 3.0: Being Human in the Age of Artificial Intelligence*, Brockman Inc, New York.

Yudkowsky, E. (2008) Artificial intelligence as a positive and negative factor in global risk, in *Global Catastrophic Risks* (eds Bostrom, N. et Ćirković, M.), Oxford University Press, Oxford, pp 308-345.

Yudkowsky, E. (2013) *Intelligence Explosion Microeconomics*, Machine Intelligence Research Institute, Berkeley, CA.

Yudkowsky, E. (2017) *There's No Fire Alarm for Artificial General Intelligence*, Machine Intelligence Research Institute, Berkeley, CA.

6. Vers une charte mondiale de l'intelligence artificielle

Thomas Metzinger

Introduction

L'heure est venue de poursuivre le débat public en cours sur l'intelligence artificielle (IA) au sein même des institutions politiques. De nombreux experts pensent que nous approchons d'un tournant de l'histoire au cours de la décennie à venir, et que la fenêtre temporelle permettant de définir l'éthique appliquée de l'IA va se refermer. Les institutions politiques doivent par conséquent définir *et* mettre en œuvre un ensemble minimal mais suffisant de règles éthiques et juridiques pour l'utilisation bénéfique et le développement futur de l'IA. Elles doivent également mettre sur pied un processus de discussion critique rationnel et fondé sur des données factuelles visant à mettre à jour, à améliorer et à réviser en permanence le premier ensemble de contraintes normatives. Vu la situation actuelle, l'issue par défaut est que les valeurs qui sous-tendent le développement de l'IA seront définies par un très petit nombre de personnes, par de grandes entreprises privées et par les institutions militaires. Il convient donc de s'atteler dès aujourd'hui à intégrer le plus grand nombre possible de perspectives.

De nombreuses initiatives différentes ont déjà été lancées dans le monde entier et étudient activement les avancées récentes dans l'IA du point de vue de l'éthique appliquée, de ses aspects juridiques, des futures conséquences socioculturelles , des risques existentiels et de l'élaboration des politiques.[4]Le débat public est animé, et certains pourraient même penser que les grandes institutions politiques telles que l'Union européenne ne sont pas en mesure de réagir suffisamment vite aux nouveaux risques technologiques et à l'inquiétude croissante de l'opinion publique. Il convient donc d'accroître la souplesse, l'efficacité et le caractère systématique des efforts politiques en cours visant à appliquer des règles au moyen de l'élaboration d'un processus démocratique plus formel et institutionnalisé, et peut-être même de nouveaux modèles de gouvernance.

Afin de lancer un processus plus systématique et structuré, je présenterai une liste concise et non exclusive des cinq problématiques les plus importantes, assorties dans chaque cas de recommandations pratiques. La première problématique à examiner est celle qui, selon moi, regroupe les problèmes qui ont le moins de chances d'être résolus. Il convient donc de l'aborder selon un processus à plusieurs niveaux, en commençant au niveau de l'Union européenne (UE).

Le problème du «nivellement par le bas»

Nous devons élaborer et mettre en œuvre pour la recherche en IA des normes de sécurité à l'échelle mondiale. Une charte *mondiale* de l'IA est nécessaire étant donné que ces normes ne pourront être efficaces que si elle s'accompagne l'engagement contraignant en faveur de certaines règles de *tous* les pays participants qui investissent dans les activités de recherche et développement concernées. Compte tenu du contexte actuel de compétitivité économique et militaire , la sécurité de la recherche en IA sera très probablement revue à la baisse pour accélérer les progrès et réduire les coûts, notamment par un transfert vers des pays appliquant des normes de sécurité moins strictes et n'offrant que peu de transparence politique (une analogie évidente et profonde peut être établie avec le phénomène de la fraude fiscale par les entreprises et les fiducies). En cas de coopération et de coordination internationales fructueuses, il serait en principe possible d'éviter un nivellement par le bas des normes de sécurité (par

[4] Pour un aperçu des initiatives en cours, voir par exemple Baum 2017 et Boddington 2017, p 3p. Je me suis abstenu de dresser ici une liste complète de références, mais le lecteur pourra découvrir la littérature à ce sujet grâce à certains documents intéressants comme Mannino *et al.* 2015, Stone *et al.* 2016, IEEE 2017, Bostrom, Dafoe & Flynn 2017, Madary & Metzinger 2016 (pour la réalité virtuelle).

la délocalisation des recherches scientifiques et industrielles en matière d'IA). Cependant, les mesures d'incitation actuelles font que cette issue est très peu probable.

Recommandation 1

L'Union devrait élaborer immédiatement une charte européenne de l'IA.

Recommandation 2

Parallèlement, l'Union devrait lancer un processus politique menant à l'élaboration d'une charte mondiale de l'IA.

Recommandation 3

L'Union devrait consacrer des ressources au renforcement systématique de la coopération et de la coordination internationales. Il convient de réduire le plus possible la méfiance stratégique et de définir des points communs en échafaudant les scénarios les plus pessimistes.

La deuxième problématique à examiner constitue peut-être l'ensemble le plus pressant de problèmes, et ces problèmes ont eux aussi assez peu de chances d'être résolus.

Prévention d'une course aux armements intelligents

Les citoyens de l'Union ont intérêt à ce qu'une course aux armements dotés de l'IA, par exemple entre la Chine et les États-Unis, soit empêchée à un stade précoce. Encore une fois, il est peut-être déjà trop tard, et l'influence de l'Europe est évidemment limitée, mais il convient d'interdire et de ne pas financer sur le territoire de l'Union les recherches et le développement d'armements autonomes offensifs. Les armes autonomes choisissent et attaquent leurs cibles sans intervention humaine. Elles vont agir de plus en plus rapidement, de sorte qu'il sera rationnel de confier une partie toujours plus importante de l'autonomie humaine à ces systèmes. Il peut en résulter des contextes militaires dans lesquels il sera rationnel d'abandonner presque entièrement le contrôle humain. Cette problématique est encore plus complexe que la prévention du développement et de la prolifération des armes nucléaires, par exemple, parce que la plupart des recherches en la matière ne sont pas réalisées au sein d'universités publiques. En outre, si l'humanité se lance dans une course aux armements à ce nouveau niveau technologique, le processus historique d'une course aux armements pourrait devenir lui-même autonome et résister aux interventions du monde politique.

Recommandation 4

L'Union devrait interdire *toutes* les recherches portant sur des armes autonomes offensives sur son territoire et chercher à conclure des accords internationaux.

Recommandation 5

Pour les applications militaires purement défensives, l'Union devrait financer des recherches sur le degré maximal d'autonomie des systèmes intelligents qui semble acceptable d'un point de vue éthique et juridique.

Recommandation 6

Sur le plan international, l'Union devrait lancer une initiative de grande ampleur visant à empêcher une course aux armements dotés d'IA en recourant à tous les moyens diplomatiques et politiques disponibles.

La troisième problématique à examiner est celle dont l'horizon de prédiction est probablement encore assez lointain, mais qui présente une incertitude épistémique élevée et qui pourrait avoir des conséquences néfastes extrêmes.

Un moratoire sur la phénoménologie synthétique

Il est important que tous les responsables politiques comprennent la différence entre intelligence artificielle et conscience artificielle. La création intentionnelle ou non d'une conscience artificielle est hautement problématique du point de vue éthique parce qu'elle pourrait engendrer une souffrance artificielle et une conscience de soi chez les systèmes intelligents autonomes. Le terme «phénoménologie synthétique» (PS, par analogie avec la «biologie synthétique») désigne la possibilité de créer non seulement une intelligence générale, mais aussi un état de conscience ou des expériences subjectives chez les systèmes artificiels avancés. Les sujets d'expérience artificiels futurs n'ont aucune représentation dans le processus politique actuel, ils n'ont pas de statut juridique et leurs intérêts ne sont représentés dans aucun comité d'éthique. Pour prendre des décisions éthiques, il est important de comprendre les systèmes naturels et artificiels qui ont la capacité de produire une conscience, et notamment de connaître des états négatifs tels que la souffrance[5]. Un risque potentiel est d'augmenter considérablement la souffrance générale dans l'univers, par exemple par des copies en cascade ou par la reproduction rapide de systèmes conscients à grande échelle.

Recommandation 7

L'Union devrait interdire toute recherche présentant un risque ou ayant pour objectif direct de développer une phénoménologie synthétique sur son territoire, et chercher à conclure des accords internationaux.[6]

Recommandation 8

Compte tenu du degré actuel d'incertitude et de désaccord dans le domaine naissant de la conscience des machines, il est urgent de promouvoir, de financer et de coordonner des projets de recherche interdisciplinaires pertinents (dont la philosophie, la neuroscience et l'informatique). Les thèmes pertinents spécifiques sont notamment les modèles conceptuels, neurobiologiques et informatiques d'expérience consciente, de conscience de soi et de souffrance fondés sur des données factuelles.

Recommandation 9

Au niveau de la recherche fondamentale, il est nécessaire de promouvoir, de financer et de coordonner des recherches systématiques sur l'éthique appliquée de systèmes non biologiques capables d'expérience consciente, de conscience de soi et de souffrance ressentie sur le plan subjectif.

La problématique générale suivante à examiner est la plus complexe et celle qui englobe probablement le plus grand nombre de problèmes inattendus et d'«inconnues inconnues».

Dangers pour la cohésion sociale

Une technologie d'IA avancée offrira manifestement de nombreuses possibilités d'optimiser le processus politique, y compris de nouvelles possibilités d'ingénierie sociale rationnelle et fondée sur des valeurs ainsi que la possibilité de créer des formes de gouvernance plus efficaces fondées sur des données factuelles. Par ailleurs, il est non seulement plausible de supposer que de nombreux nouveaux risques et dangers inconnus à ce jour sont susceptibles de saper la cohésion de nos sociétés, mais il est aussi rationnel de supposer l'existence d'un nombre plus important d'«inconnues inconnues», de risques liés à l'IA que nous ne découvrirons que par hasard et à un stade ultérieur. L'Union devrait donc

[5] Voir Metzinger 2013, 2017.

[6] Ces recherches incluent les approches visant une convergence de la neuroscience et de l'IA dans le but spécifique de développer une conscience chez les machines. Pour des exemples récents, voir Dehaene, Lau & Kouider 2017, Graziano 2017, Kanai 2017.

allouer des *ressources distinctes* pour se préparer à des situations dans lesquelles nous découvrirons soudainement de telles «inconnues inconnues» inattendues.

De nombreux experts pensent que le risque le plus immédiat et le mieux défini est celui d'un chômage de masse résultant de l'automatisation. La mise en œuvre de la technologie de l'IA par des parties prenantes financièrement puissantes pourrait donc entraîner un accroissement des inégalités de revenus et d'autres inégalités ainsi que des schémas dangereux de stratification sociale. Concrètement, on pourrait s'attendre à une forte baisse des salaires, à un effondrement des recettes fiscales et à une sollicitation excessive des systèmes de sécurité sociale. L'IA pose néanmoins de nombreux autres risques pour la cohésion sociale, par exemple par des médias sociaux privés et autonomes visant à solliciter l'attention du public et à l'«emballer» pour une utilisation ultérieure par leurs clients, ou encore par la «création de toutes pièces» d'une volonté politique au moyen des stratégies de «suggestion» à grande échelle et des architectures de choix contrôlées par l'IA et opaques pour les citoyens dont elles contrôlent le comportement. La future technologie de l'IA sera extrêmement efficace dans la modélisation et le contrôle prédictif du comportement humain, par exemple par un renforcement positif et des suggestions indirectes, de sorte que lerespect de certaines normes ou l'apparition «spontanée» de «motifs» et de décisions semblera se faire sans aucune contrainte. Associée aux suggestions à grande échelle et au contrôle prédictif des utilisateurs, la technologie de surveillance intelligence pourrait également accroître les risques mondiaux en contribuant *localement* à stabiliser efficacement les régimes autoritaires. Ici aussi, il est très probable que la plupart des risques pour la cohésion sociale soient encore inconnus à l'heure actuelle, et il se peut que nous ne les découvrions que par hasard. Les décideurs politiques doivent également comprendre que toute technologie capable d'optimiser intentionnellement l'intelligibilité de ses propres actions pour les utilisateurs humains peut aussi, en principe, optimiser la *tromperie*. Il convient donc d'éviter soigneusement de définir (intentionnellement ou non) la fonction de récompense d'une IA d'une façon susceptible de nuire de manière indirecte au bien commun.

La technologie d'IA constitue actuellement un bien privé. Les institutions politiques démocratiques ont le devoir de transformer une grande partie de cette technologie en un bien *commun* bien protégé, qui appartienne à l'humanité dans son ensemble. Dans la tragédie des biens communs, tout le monde voit souvent venir la catastrophe mais, si aucun mécanisme n'est en place pour l'éviter, par exemple dans les situations décentralisées, cette catastrophe se produira malgré tout. L'Union devrait sans plus attendre s'atteler à mettre de tels mécanismes en place.

Recommandation 10

Au sein de l'Union, les gains de productivité découlant de l'IA doivent être distribués de manière socialement équitable. Les pratiques antérieures et les tendances mondiales vont clairement dans le sens opposé: nous n'avons (presque) jamais procédé de la sorte dans le passé, et les mesures d'incitation financières actuelles contredisent directement cette recommandation.

Recommandation 11

L'Union devrait étudier soigneusement la possibilité d'un revenu de base inconditionnel ou d'un impôt négatif sur le revenu sur son territoire.

Recommandation 12

Il convient de mettre sur pied des programmes de recherche sur la possibilité de lancer le moment voulu des initiatives de reconversion aux compétences créatives et sociales des couches de la population qui se voient menacées.

La problématique suivante est sensible car la majeure partie des recherches de pointe en IA est déjà sortie du cadre des universités et des établissement de recherche bénéficiant d'un financement public.

Ces recherches sont désormais aux mains d'entreprises privées, et manquent donc systématiquement de transparence.

Éthique de la recherche

L'un des problèmes théoriques les plus difficiles consiste à définir les conditions dans lesquelles il serait rationnel d'abandonner entièrement certains types de recherche en matière d'IA (par exemple, les recherches sur l'apparition d'une phénoménologie synthétique, ou sur l'évolution exponentielle de systèmes qui s'optimisent de manière autonome sans se conformer de manière fiable aux valeurs humaines). Quels scénarios concrets et minimaux justifieraient un moratoire sur certains types de recherches? Comment les institutions démocratiques pourront-elles faire face à des acteurs délibérément non éthiques, lorsqu'une prise de décisions collective n'est pas réaliste et qu'il convient de mettre en place des formes graduelles et non globales de coopération ad hoc? Des problèmes similaires se sont déjà posés dans la «recherche sur les gains de fonction», avec une expérimentation visant à accroître la transmissibilité et/ou la virulence de pathogènes, comme certaines souches hautement pathogènes du virus de la grippe H5N1, de la variole ou de la maladie du charbon. Dans ce contexte, et tout à leur honneur, les chercheurs sur la grippe se sont imposé un moratoire volontaire et temporaire. Une telle approche serait en principe possible dans la communauté des chercheurs en matière IA. L'Union devrait donc veiller à toujours compléter sa charte de l'IA par un code de déontologie concret pour les chercheurs qui participent à des projets financés par l'Union.

L'objectif plus profond devrait cependant être de développer une *culture de sensibilité morale* plus globale au sein des communautés de chercheurs concernées. Toute recherche devrait inclure des mesures d'identification et de réduction des risques rationnelles et fondées sur des données factuelles (y compris pour les risques à prévoir dans un avenir lointain), et les scientifiques devraient se montrer vigilants et chercher à anticiper, en particulier lorsqu'ils sont les premiers à prendre conscience de nouveaux risques grâce à leurs travaux. La communication avec le public, si nécessaire, doit se faire spontanément. Il s'agit de prendre le contrôle et d'agir avant l'apparition d'une situation future plutôt que de réagir aux critiques de personnes non spécialisées par un ensemble de règles formelles préexistantes. Comme Madary et Metzinger (2016, p. 12) l'écrivent dans leur code déontologique, qui comprend des recommandations de bonnes pratiques scientifiques en réalité virtuelle: «Les scientifiques doivent comprendre qu'il existe une différence entre respecter un code de déontologie et *faire preuve soi-même de déontologie*. Un code de déontologie propre à un domaine, aussi cohérent, avancé et précis qu'il puisse être dans ses versions futures, ne peut jamais se substituer à un raisonnement inspiré par la déontologie proprement dite.»

Recommandation 13

Toute charte mondiale de l'IA, ou son précurseur européen, devrait toujours être accompagnée d'un code de déontologie guidant les chercheurs dans leurs travaux pratiques au quotidien.

Recommandation 14

Il convient de former une nouvelle génération d'éthiciens spécialisés dans les problèmes liés à l'IA, aux systèmes autonomes et aux domaines connexes. L'Union devrait investir sans délai et en bon ordre dans le développement de l'expertise future nécessaire au sein des institutions politiques concernées, tout en aspirant à un niveau d'excellence académique et de professionnalisme supérieur à la moyenne et particulièrement élevé.

La métagouvernance et l'écart de vitesse

Comme évoqué brièvement dans le paragraphe introductif, l'accélération du développement de l'IA est sans doute devenu l'exemple *paradigmatique* d'un décalage extrême entre les approches existantes des pouvoirs publics et ce qui serait nécessaire pour optimiser en temps opportun le rapport

risque/avantages. Il s'agit d'un exemple paradigmatique de contraintes temporelles en ce qui concerne l'identification, l'évaluation et la gestion rationnelles et fondées sur des données factuelles des risques émergents, l'élaboration de lignes directrices éthiques et la mise en œuvre d'un ensemble de règles juridiques applicables. Il existe un «problème de rythme»: les structures de gouvernance existantes sont tout simplement incapables de réagir assez vite à ce problème. Le contrôle politique est déjà largement à la traîne de l'évolution technologique[7].

Je ne cherche pas à attirer l'attention sur la situation actuelle dans le but de me montrer alarmiste ou de conclure sur une note dystopique et pessimiste. Je tiens plutôt à montrer que l'adaptation des structures de gouvernance *elles-mêmes* s'inscrit dans cette problématique: pour combler ou au moins réduire l'écart entre ces différents rythmes d'évolution, nous devons consacrer des ressources à la modification de nos propres approches des structures de gouvernance. C'est exactement ce que désigne le terme de «métagouvernance»: une gouvernance de la gouvernance face aux risques et aux avantages potentiels d'une croissance exponentielle dans des secteurs spécifiques de développement technologique. Wendel Wallach, par exemple, a observé que le contrôle efficace des technologies émergentes nécessitait une combinaison de réglementations «dures» imposées par les pouvoirs publics et de mécanismes de gouvernance informelle élargis.[8]C'est pourquoi Marchant et Wallach ont proposé des «Communautés de coordination de la gouvernance» (CCG), un nouveau type d'institution créant un mécanisme de coordination et de synchronisation de ce qu'ils décrivent à juste titre comme une «explosion de stratégies, d'actions, de propositions et d'institutions de gouvernance»[9] avec les travaux existants au sein des institutions politiques établies. Une CCG pour l'IA pourrait faire office de «gestionnaire de problèmes» pour une technologie précise en proie à une émergence rapide, de centre de documentation, de système d'alerte précoce, d'instrument d'analyse et de suivi, d'évaluateur international des bonnes pratiques et de référence indépendante et de confiance pour les éthiciens, les médias, les scientifiques et les parties prenantes concernées. Comme l'écrivent Marchant et Wallach: «*L'influence d'une CCG pour répondre au besoin critique d'une entité de coordination centrale dépendra de sa capacité à s'imposer comme un médiateur impartial respecté par toutes les parties concernées.*»[10]

On peut évidemment envisager de nombreuses autres stratégies et approches de la gouvernance, mais elles ne seront pas détaillées dans le présent article. Sa conclusion générale est tout simplement que nous ne pouvons relever le défi posé par l'évolution rapide de l'IA et des systèmes autonomes que si nous faisons de la métagouvernance la question centrale de nos réflexions.

Recommandation 15

L'Union devrait investir dans la recherche et dans le développement de nouvelles structures de gouvernance permettant aux institutions politiques établies de réagir aux problèmes et appliquer concrètement de nouvelles réglementations bien plus rapidement qu'aujourd'hui.

---

[7] Gary Marchant (2011) exprime très clairement ce problème général dans le résumé d'un chapitre de livre écrit récemment: «*Les technologies émergentes se développent de plus en plus vite, tandis que les mécanismes juridiques de contrôle potentiel ont plutôt tendance à ralentir. La législation se retrouve souvent dans l'impasse, la réglementation est souvent figée et les procédures judiciaires avancent parfois à un rythme d'escargot. Ce décalage entre rythme de l'évolution de la technologie et celui de la loi a deux conséquences. Tout d'abord, certains problèmes sont soumis à des cadres réglementaires de plus en plus obsolètes et dépassés. Ensuite, d'autres problèmes ne font l'objet d'aucun contrôle digne de ce nom. Pour combler cet écart croissant entre la loi et la réglementation, de nouveaux outils, approches et mécanismes juridiques seront nécessaires. Poursuivre dans la voie actuelle ne suffira pas.*»

[8] Voir Wallach 2015 (Chapitre 14), p. 250.

[9] Cette situation est issue d'une version préliminaire non publiée intitulée «*An agile ethical/legal model for the international and national governance of AI and robotics*»; voir aussi Marchant & Wallach 2015.

[10] Marchant & Wallach 2015, p. 47.

Conclusions

J'ai proposé que l'Union s'attèle sans plus attendre à élaborer une charte mondiale de l'IA, selon un processus en plusieurs étapes commençant par une charte de l'IA pour l'Union européenne elle-même. Pour illustrer brièvement certaines des questions fondamentales depuis ma propre perspective de philosophe, j'ai recensé cinq problématiques principales et formulé quinze recommandations générales en vue d'un débat critique. Il va de soi que cette contribution ne prétend pas constituer une liste exclusive ou exhaustive des questions pertinentes. Au contraire: à l'origine, l'éthique appliquée de l'IA ne se prête pas du tout aux grandes théories ni aux débats idéologiques, c'est plutôt une question de gestion réfléchie et rationnelle des risques selon plusieurs horizons de prévision et moyennant un degré important d'incertitude. Un élément important du problème est toutefois que nous ne pouvons pas nous fier à nos intuitions, parce que nous devons respecter des contraintes de rationalité contre-intuitive.

Permettez-moi de conclure en citant un document d'orientation récent intitulé *Artificial Intelligence: Opportunities and Risks*, publié par l'*Effective Altruism Foundation* de Berlin, Allemagne:

Dans les situations de prise de décision aux enjeux cruciaux, les principes suivants sont d'une importance primordiale:

1. des précautions coûteuses peuvent justifier ces coûts même pour les risques à faible probabilité, pour autant que les gains/pertes potentiels soient suffisants.
2. Faute de consensus entre les experts, la modestie épistémique est préférable: mieux vaut éviter d'avoir trop confiance dans l'exactitude de ses propres avis, dans un sens ou dans l'autre[11].

Références

Adriano, Mannino; Althaus, David; Erhardt, Jonathan; Gloor, Lukas; Hutter, Adrian; Metzinger, Thomas (2015): Artificial Intelligence. Opportunities and Risks. In: *Policy Papers of the Effective Altruism Foundation* (2), S. 1–16. https://ea-foundation.org/files/ai-opportunities-and-risks.pdf.

Baum, Seth (2017): A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy. Global Catastrophic Risk Institute Working Paper 17-1. https://ssrn.com/abstract=3070741.

Boddington, Paula (2017): Towards a Code of Ethics for Artificial Intelligence. Cham Springer International Publishing (Artificial Intelligence: Foundations, Theory, and Algorithms).

Bostrom, Nick; Dafoe, Allan; Flynn, Carrick (2017): Policy Desiderata in the Development of Machine Superintelligence. working Paper, Oxford University. http://www.nickbostrom.com/papers/aipolicy.pdf.

Dehaene, Stanislas; Lau, Hakwan; Kouider, Sid (2017): What is consciousness, and could machines have it? In: *Science (New York, N.Y.)* 358 (6362), S. 486–492. DOI: 10.1126/science.aan8871.

Graziano, Michael S. A. (2017): The Attention Schema Theory. A Foundation for Engineering Artificial Consciousness. In: *Frontiers in Robotics and AI* 4, S. 61. DOI: 10.3389/frobt.2017.00060.

Madary, Michael; Metzinger, Thomas K. (2016): Real virtuality. A code of ethical conduct. recommendations for good scientific practice and the consumers of VR-technology. In: *Frontiers in Robotics and AI* 3, S. 3. http://journal.frontiersin.org/article/10.3389/frobt.2016.00003/full

Marchant, Gary E. (2011): The growing gap between emerging technologies and the law. In Marchant, Gary E.; Allenby, Braden R.; Herkert, Joseph R. (Hg.): The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: Springer, S. 19–33.

[11] Cf. Mannino *et al*. 2015.

Marchant, Gary E.; Wallach, Wendell (2015): Coordinating technology governance. In: *Issues in Science and Technology* 31 (4), S. 43.

Metzinger, Thomas (2013): Two principles for robot ethics. In: In Hilgendorf, Eric; Günther, Jan-Philipp (Hg.) (2013): Robotik und Gesetzgebung: BadenBaden, Nomos S. 247–286. https://www.blogs.uni-mainz.de/fb05philosophieengl/files/2013/07/Metzinger_RG_2013_penultimate.pdf

Metzinger, Thomas (2017): Suffering. In: Kurt Almqvist und Anders Haag (Hg.): The Return of Consciousness. Stockholm: Axel and Margaret Ax:son Johnson Foundation, S. 237–262. https://www.blogs.uni-mainz.de/fb05philosophieengl/files/2013/07/Metzinger_Suffering_2017.pdf

Kanai, Ryota (2017): We Need Conscious Robots. How introspection and imagination make robots better. In: *Nautilus* (47). http://nautil.us/issue/47/consciousness/we-need-conscious-robots.

Stone, Peter; et al. (2016): Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel. Stanford, CA: Stanford University. https://ai100.stanford.edu/2016-report.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2017): Ethically Aligned Design. A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. http://standards.ieee.org/develop/indconn/ec/auto_sys_form.html.

Wallach, W. (2015): A Dangerous Master. How to Keep Technology from Slipping Beyond Our Control. New York: Basic Books.

Pour le meilleur ou pour le pire, l'intelligence artificielle (IA) aura probablement une incidence considérable sur l'avenir de l'humanité. Les nouvelles promesses et inquiétudes trouvent un écho de plus en plus large auprès du grand public, et le débat commence à capter son imagination. Dans cette publication, nous présentons quatre articles d'opinion qui répondent tous à la même question: faut-il craindre l'intelligence artificielle? Les quatre auteurs sont issus de disciplines différentes et présentent des perspectives divergentes, illustrant les raisons pour lesquelles il convient de craindre ou non l'avenir de l'IA et comment gérer son évolution.

> *Les progrès de l'intelligence artificielle ont suscité d'énormes espoirs et suscité des craintes tout aussi importantes, souvent dénuées de fondement dans la réalité. Ce formidable recueil de textes écrits par de réels experts propose une analyse rationnelle de ce sujet chargé d'émotions et constituera un outil indispensable pour quiconque souhaite comprendre l'un des enjeux les plus importants de notre époque.*
>
> Steven Pinker Johnstone, Professeur de psychologie, université d'Harvard et auteur de *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress.*

Étude publiée par l'Unité de la prospective scientifique (STOA)
EPRS | Service de recherche du Parlement européen

PE 614.547
ISBN 978-92-846-3387-6
doi: 10.2861/587509
QA-01-18-199-FR-N